全国统计教材编审委员会“十二五”规划教材

统计学：从数据到结论

第四版

吴喜之 编著

中国统计出版社
China Statistics Press

责任编辑 梁超
封面设计 上智博文

统计学：从数据到结论 第四版

全国统计教材编审委员会“十二五”规划教材

ISBN 978-7-5037-6789-0

定价：30.00元

全国统计教材编审委员会“十二五”规划教材

统计学：从数据到结论

第四版

吴喜之 编著

中国统计出版社
China Statistics Press

图书在版编目(CIP)数据

统计学：从数据到结论 / 吴喜之编著. —4 版. — 北京 ：中国统计出版社，2013.3

全国统计教材编审委员会“十二五”规划教材

ISBN 978—7—5037—6789—0

Ⅰ. ①统… Ⅱ. ①吴… Ⅲ. ①统计学 Ⅳ. ①C8

中国版本图书馆 CIP 数据核字(2013)第 044840 号

统计学：从数据到结论

作　　者/吴喜之
责任编辑/梁　超
封面设计/上智博文
出版发行/中国统计出版社
通信地址/北京市丰台区西三环南路甲 6 号　　邮政编码/100073
电　　话/邮购(010)63376909　书店(010)68783171
网　　址/http://csp.stats.gov.cn/
印　　刷/河北天普润印刷厂
经　　销/新华书店
开　　本/710×1000mm　1/16
字　　数/267 千字
印　　张/15
印　　数/1—5000 册
版　　别/2013 年 3 月第 4 版
版　　次/2013 年 3 月第 1 次印刷
定　　价/30.00 元

版权所有。未经许可，本书的任何部分不得以任何方式在世界任何地区以任何文字翻印、拷贝、仿制或转载。

中国统计版图书，如有印装错误，本社发行部负责调换。

全国统计教材编审委员会

顾　问　罗　兰　袁　卫　冯士雍　吴喜之
方积乾　王吉利　庞　皓　李子奈

主　任　徐一帆

副主任　严建辉　田鲁生　邱　东　施建军
耿　直　徐勇勇

委　员（按姓氏笔划排序）

丁立宏　万崇华　马　骏　毛有丰　王兆军
王佐仁　王振龙　王惠文　丘京南　史代敏
龙　玲　刘建平　刘俊昌　向书坚　孙秋碧
朱　胜　朱仲义　许　鹏　余华银　张小斐
张仲梁　张忠占　李　康　李兴绪　李宝瑜
李金昌　李朝鲜　杨　虎　杨汭华　杨映霜
汪荣明　肖红叶　苏为华　陈　峰　陈相成
房祥忠　林金官　罗良清　郑　明　柯惠新
柳　青　胡太忠　贺　佳　赵彦云　赵耐青
凌　亢　唐年胜　徐天和　徐国祥　郭建华
崔恒建　傅德印　景学安　曾五一　程维虎
蒋　萍　潘　璠　颜　虹

出版说明

“十二五”时期，是我国全面实施素质教育，全面提高高等教育质量，深化教育体制改革，推动教育事业科学发展，提高教育现代化水平的时期。“十二五”伊始，统计学迎来了历史性的重大变革和飞跃。2011年2月，在国务院学位委员会第28次会议通过的新的《学位授予和人才培养学科目录(2011)》(以下简称“学科目录”)中，统计学从数学和经济学中独立出来，成为一级学科。这一变革和飞跃将对中国统计教育事业产生巨大而深远的影响，中国统计教育事业将在“十二五”时期发生积极变化。

正是在这一背景下，全国统计教材编审委员会制定了《“十二五”全国统计教材建设规划》(以下简称“规划”)。根据“学科目录”在统计学下设有数理统计学，社会经济统计学，生物卫生统计学，金融统计、风险管理与精算学，应用统计5个二级学科的构架，“规划”对“十二五”全国统计规划教材建设作了全面部署，具有以下特点：

第一，打破以往统计规划教材出版学科单一的格局。全面发展数理统计学，社会经济统计学，生物卫生统计学，金融统计、风险管理与精算学，应用统计5个二级学科规划教材的出版，使“十二五”全国统计规划教材涵盖5个二级学科，形成学科全面并平衡发展的出版局面。

第二，打破以往统计规划教材出版层次单一的格局。在编写出版好各学科本科生教材的基础上，对研究生教材出版进行深入研究，出版一批高水平高层次的研究生教材，为我国研究生教育、尤其是应用统计研究生教育提供教学服务。同时，积极重视统计专科教材出版，联合各专科院校，组织编写和出版适应统计专科教学和学习的优秀教材。

第三，打破以往统计规划教材出版品种单一的格局。鼓励内容创新，联系统计实践，具有教学内容和教学方法特色的、各高校自编的相同内容选题的精品教材出版，促进统计教学向创新性、创造性和多样性

发展。

第四，重视非统计专业的统计教材出版。探讨对非统计专业学生的统计教学问题，为非统计专业学生组织编写和出版概念准确、叙述简练、深入浅出、表达方式活泼、练习题贴近社会生活的统计教材，使统计思想和统计理念深入非统计专业学生，以达到统计教学的最大效果。

第五，重视配合教师教学使用的电子课件和辅助学生学习使用的电子产品的配套出版，促进高校统计教学电子化建设，以期最后能形成系统，提高统计教育现代化水平。

第六，重视对已经出版的统计规划教材的培育和提高，本着去粗存精、去旧加新、与时俱进的原则，继续优化已经出版的统计教材的内容和写作，强化配套课件和习题解答，使它们成为精品，最后锤炼成为经典。

“十二五”期间，编审委员会将本着“重质量，求创新，出精品，育经典”的宗旨，组织我国统计教育界专家学者，编写和编辑出版好本轮教材。本轮教材出版后，将能够形成学科齐全、层次分明、品种多样、配套系统的高质量立体式结构，使我国统计规划教材建设再上新台阶，这将对推动我国统计教育和统计教材改革，推动我国统计教育事业科学发展，提高我国统计教育现代化水平产生积极意义。

让教师的教学和学生的学习事半功倍，并使学生在毕业之后能够学以致用的统计教材，是本轮教材的追求。编审委员会将努力使本轮教材好教、好学、好用，尽力使它们在内容上和形式上都向国外先进统计教材看齐。限于水平和经验，在教材的编写和编辑出版过程中仍会有不足，恳请广大师生和社会读者提出批评和建议，我们将虚心接受，并诚挚感谢！

全国统计教材编审委员会

2012 年 7 月

再版说明

这本书已经有了近十年的历史，现在将要出第四版。前面三版已经作为参考书或教科书在许多学校使用。各个学校的师生对本书提出许多宝贵的意见，并且指出了很多错误和不妥之处。读者的支持和鼓励，对本书各版的诞生起着关键的作用。第四版在许多地方对前面几版进行了修改和增减。

免费的自由编程的R软件在国际上已经成为统计教学和科研的主要软件，本书第四版全部采用R软件来描述计算过程，彻底放弃了使用商业软件。R软件非常强大，凡是国际上出现的新方法，都会很快地上传到R的网站上，在发达国家，不能想象一个统计教师或者统计研究生不会熟练使用R。从R的功能和使用者的人数来说，它已经远远超过所有昂贵的商业软件。R软件的绝大部分程序包的代码都是公开的，透明是防止腐败的最好方式。此外，由于R在中国的普及越来越广泛，网上关于R的互动和帮助的环境也已经形成，中国学生和实际工作者完全可以赶上国际统计界使用R的主流(虽然已经至少落后了10年)。

在强大的免费R软件不断普及的情况下，对于缺乏经费的中国教育系统以及并非富裕的学校师生来说，教学中继续通过昂贵的商业软件来讲授统计变得越来越缺乏吸引力。用商业软件教学的一个客观效果是鼓励非法盗版行为。由于避免了对商业软件菜单的点击鼠标的繁琐而又冗长的细节叙述，整本书都显得简洁明了，节省了大量的篇幅(第四版比前一版减少了一百多页)。课文中所有计算过程都附有可以实现的R语句，在每章最后仅仅对R语句做些汇总或说明。

虽然R软件是编程语言，但由于其简单易懂，任何从来没有使用过R的人都可以毫不费力地通过复制和粘贴书上的代码重新实现书上的所有例题。书后附录中的R代码练习更可以帮助读者尽快地掌握R语言。

许多人，比如各层管理人员，并不一定都进行第一线的实际数据计算，但为了理解手中关于本单位及有关方面信息的意义，为了更好地进行明白的决策，他们必须理解各种统计推断结果的意义。对这些人，不一定要求能够使用软件，更不需要理解数学推导，但他们必须明白各种统计概念和方法以及输出结果的意义，明白那些数据分析人员在做什么。相信本书对他们肯定会有所裨益。

在内容方面，本版专门添加了有广泛应用前景的机器学习的回归和分类方法，并且把这些内容及经典的回归和判别分析等归到一章。此外，把多元分析的除判别分析之外的其他内容合并到一章之中。这一版还取消了非参数检验一章，把其中一些常用的非参数检验加入到假设检验的一章中。

作为教科书，本书内容对于每周两学时的课程似乎太多。我觉得，什么讲或者什么不讲应该根据学生的需要由老师自己安排。实际上，对于任何课程，最好是由任课教师来决定讲哪些内容以及如何讲。因为他们最了解他们所面对的学生。教科书编者的思维方式不见得和老师的一致，而老师最好按照自己的理解来讲述。一个好的教科书，应该给教师以较大的余地和自由。

笔者希望读者在阅读本书时能够以理解统计方法的含义为主，学会处理数据，提高学习和应用能力。**在任何国家及任何制度下都能够生存和发展的知识和能力，就是科学，是人们在生命的历程中应该获得的。**

希望读者继续对本书予以宝贵的支持和批评指正。

吴喜之

2012 年 10 月

第一版前言

什么在本书中等待着你们去发现，去探讨，去欣赏呢？当然不是数学公式和定理定义的堆砌，也不是和枯燥的公文报表相关的政府工作的培训。这是一门充满了哲学韵味的认识世界的学问。

不知读者们是否意识到，统计已经渗入到人们的社会、生活、工作等各个领域。每天新闻媒介报道的各个方面都离不开各种统计数据和各种分析与预测。人们可能对于这些统计内容觉得习以为常，也可能会有一些好奇或神秘感。由于国情不同，统计的地位与人们对统计的看法也不同。在发达国家，一般民众觉得统计学和数学类似，是一门高不可攀但极易找到满意工作的学问。在中国，又有一些人认为统计就是处理政府报表的职业。但自从中国向世界开放之后，越来越明确的一点是，没有什么学科或领域能够真正离开统计。

以应用为目标学习统计，究竟是为什么？是为了流利地背诵一大堆定义、概念和抽象的名词和术语吗？是为了学习如何进行推导和证明一些复杂的定理和公式吗？这些问题不仅学生会思考，更重要的是统计教师要思考。本书的目的是希望读者在学习之后，能够知道实际中哪些是统计问题，最好能够自己解决一部分统计问题，即使不能解决也知道能够在哪里查到答案和向谁请教。知识固然重要，更重要的是通过学习获得解决和处理问题的能力。

学习并不总是一个令人生畏或至少成为某种负担的过程。人们学会走路、说话、骑车、下棋、打球等大都是在一种乐趣中进行的。为什么涉及日常生活的每一个方面的统计就不能和看侦探小说那么引人入胜呢？其实任何一门科学，都有其趣味性，而只有把科学研究当成游戏的人才会真正成为大师。这门课并不想使读者都成为统计学家，而仅仅想让读者如同学会使用电脑、手机、学会辩论、上网或讨价还价那样愉快地认识或理解在人生中无法躲开的统计。

本书由浅入深地把统计最基本和最有用的部分在这么一本不厚的教科书中完整地介绍给读者，而且让读者可以边学习，边着手用统计软件处理数据。篇幅大、语言罗嗦的教材对读者是个负担，不但浪费了资源，也抓不住要领。因此，作者力图惜墨如金，既节省篇幅，又要把该解释的全部说清。希望读者慢慢咀嚼，不必图快。

很少有一本统计教材包括像本书那么多的统计内容。我觉得，这些内容本来并不深奥，只是其貌似复杂的数学工具把它搞成阳春白雪，再加上强调数学推导的教学方式，使得统计显得高不可攀。本教材要还这些统计应用以其本来面目。使得统计变成人人都能够基本上理解和掌握的有用工具。多数使用计算机的人都不是计算机专业的，多数开汽车的都不会修汽车，但这对他们毫无妨碍。难道不会推导或背诵与统计有关的数学公式就不能应用统计这个工具了吗？

本书每一章的主要部分是用日常语言来引进和解释一些概念，如果可能，就通过例子来说明。如果不涉及应用，这部分就足够了。在本书例题的分析中，同时提供简洁明了的软件代码，可以使读者一边看书，一边自己计算，这会给多数想要自己动手分析数据的读者以方便。每章后面的小结中还展示了与概念及计算有关的一些数学公式以及软件的说明，使那些精力充沛的读者能更深刻地理解内容。这种安排使得本教材能够适用于各种不同水平、不同要求的读者群体。

本教材不仅可供没有学过概率论和数理统计的非统计专业的本科生和研究生使用，也可以供统计专业的本科生作为理解统计本来含义的教材使用（以代替不能满足需要的“描述统计学”等类课程），它还可以为各领域的广大实际工作者作为应用各种统计方法的参考书。为了读者可以使用各种软件来进行分析，本书所涉及的所有电子版数据都为文本格式。

软件方面，本书则采用免费的自由软件 R[①]。经验表明，在学习统计内容的时候学习软件比上专门的软件操作课更有效。R 软件既采用了最简单的编程语言，又拥有最丰富的统计资源。一个大学本科生通常可以

① 第一版主要使用 SPSS 和部分地应用 Excel，第二版之后加了 SAS 和 R 的应用，第三版则以 R 软件为主，第四版则全部用 R 软件。

在一天内学会R的基本计算，在一周内学会统计基本课程的计算。

在前计算机时代，几乎所有的统计教科书都给出了各种与分布有关的表格。但随着计算机的普及，所有统计软件（无论是商业的还是免费的）都给出了和各种分布有关的各种函数，把人们从繁琐而又不精确的查表中解放出来。目前很多国外的统计教科书都不再提供既占用篇幅又比较粗糙的分布表。本书不准备提供任何和分布有关的表格。本书第四章会介绍如何使用软件来进行与概率分布有关的计算。

这个教材的全部内容曾作为非统计专业硕士和博士的课程分别在北京大学光华管理学院及中国人民大学讲授过，受到普遍欢迎。实践证明，这本书的大部分内容完全能够轻轻松松地在一个学期（每周三个学时）中全部讲完。一些热心而又好奇的非统计背景的人士也曾读过本教材的全部内容，没有任何理解上的问题。当然，根据不同的教学对象和需要，有些章节可以完全不讲或少讲。

本书前面的章节，是对统计基本概念的介绍。而后面的部分则是更有针对性的一些统计模型和方法。一般传统统计学的课程包括前六章，或最多前七章的内容，而第八章属于多元统计分析的课程内容，第九章一般属于时间序列课程包含的内容，第十章简单介绍了生存分析，第十一章对指数进行了必要的介绍。目前大多数流行的统计应用都已包含在本教材内。

本书的编写是在国家统计局教育中心的建议和鼓励下产生，并得到其大力支持。本书还受到北京大学、中国人民大学以及各兄弟院校老师和学生的鼓励和帮助。中国统计出版社一直关心着本书的写作和出版。特别要指出的是敬爱的汪仁官老师又一次为我所写的统计教材进行了非常认真的审校，使我重新感受到做学生的幸福，中国统计界的老前辈茆诗松老师也热心地对本书提出了许多宝贵而又中肯的建议。他们的审校和建议使本书避免了许多错误和不妥之处。没有这些支持和帮助，本书是不可能面世的。谨在此对所有各方面表示衷心的感谢。

吴喜之

2003年6月

目　录

第一章　一些基本概念

1.1　统计是什么?

你想过下面的问题吗?

1. 当你买了一台电脑时, 被告知三年内可以免费保修. 那么, 厂家凭什么这样说? 说多了, 厂家会损失, 说少了, 会失去竞争力, 也是损失. 到底这个保修期是怎样决定的呢?

2. 在同一年级中, 同样统计学的课程可能由一些不同教师讲授. 教师讲课方式当然不一样, 考试题目也不一定相同. 那么如何比较不同班级的统计学成绩呢?

3. 大学或企业的排名是一个非常敏感的问题. 不同的机构得出不同的结果, 各自都说自己是客观、公正和有道理的. 到底如何理解这些不同的结果呢?

4. 任何公司和个人都有一个信用问题. 如果他们在试图得到贷款时并没有不还贷的不良记录, 如何根据其背景资料来判断其信用等级呢?

5. 我国东部和西部的概念是一个比较笼统的概念. 如何能够根据某些标准或需要, 选择一些指标来把各省, 或各市县甚至村进行分类呢?

6. 疾病传播时, 如何能够通过被感染者入院前后的各种经历得到一个疾病传染方式的模型呢?

7. 如何通过问卷调查来得到性别、年龄、职业、收入等各种因素与公众对某项事物(比如商品或政策)的态度的关系呢?

8. 一个从来没有研究过红楼梦的统计学家如何根据比较写作习惯得出红楼梦从哪一段开始就不是曹雪芹的手笔了呢?

9. 如何才能够客观地得到某个电视节目的收视率, 以确定插播的广告价格是否合理呢?

10. 如何根据税务部门过去的税收记录来预测下一年的税收收入, 供政府部门制定预算时参考?

11. 如何根据某地区的寿命记录来确定人寿保险的既有竞争力, 又有利可图的定价?

其实, 这些都是统计应用的例子. 这样的例子太多了, 无法一一列举. 因为统计学可以应用于几乎所有的领域, 包括精算、农业、动物学、人类学、考古学、审计学、晶体学、人口统计学、牙医学、生态学、经济计量学、教育学、选举预测和策划、工程、流行病学、金融、水产渔业研究、遗传学、地理学、地质学、

历史研究、人类遗传学、水文学、工业、法律、语言学、文学、劳动力计划、管理科学、市场营销学、医学诊断、气象学、军事科学、核材料安全管理、眼科学、制药学、物理学、政治学、心理学、心理物理学、质量控制、宗教研究、社会学、调查抽样、分类学和气象改善、博彩、遥感、卫星数据处理、网络管理、网络数据分析等. 当然, 大家用不着也不可能理解所有的统计应用. 只要能够解决自己身边的统计问题就足够了.

在解决上面所提到的若干个应用问题时所需使用的大多数统计分析方法将会在本书后面章节中陆续介绍. 当然书中的例子并不一定就刚好是上面问题中的具体例子, 但至少所使用的分析方法是类似的.

上面的例子并没有明确说出什么是统计. 其实很简单. 上面的所有例子都要通过各种直接或间接的手段来收集**数据(data)**, 都要利用一些方法来整理和分析数据, 最后通过分析得到结论. 统计是一门科学, 它以现实世界待解决的问题为目标, 这一点, 和物理学等其他科学一样. 科学研究的方法是: 观测世界或进行试验, 得到数据, 提出可以解释这些观测的假说或理论, 试图尽可能地接近现实世界的规律, 当出现理论或假说无法解释的现象(数据)时, 就有可能需要对原有理论进行修正或者代之以新理论. 统计学的假说或理论通常称为模型. 按照不列颠百科全书关于统计的定义, **统计学(statistics)是"收集、分析、展示和解释数据的科学."** [1] 与物理学的假说类似, 统计学的模型仅仅是对现实的近似, 没有任何模型是"正确"的, 也无法证明任何模型是正确的. 只能够说, 在某些可能有争议的准则之下, 某些模型比另外一些要更合适一些. 在数学逻辑中存在的确定性在统计中完全不成立. 针对于不同学科问题而发展的统计学中的数学完全不成为一个完整封闭的体系, 也没有必要成为一个数学体系. 能否解决实际问题是评价统计方法的最终标准.

比如要得到某电视节目的收视率, 可能首先要在该节目播出时, 利用电话或别的手段对看电视的人进行采访, 同时问他们在观看什么节目. 在得到了被采访的看电视的总人数, 和其中观看该节目的人数之后, 就有可能得到这部分观众中, 观看该节目的比例, 即粗糙的收视率. 之后还要经过统计分析, 评估这个收视率的可信度和代表性等等. 显然, 这是一个收集数据, 然后通过分析数据得到结论的简单例子.

思考一下:

1. 你周围经常会有辩论, 是不是这些辩论都是以科学的方法来进行的?

2. 对世界的解释, 除了科学还有信仰, 举例说明科学和信仰之间的区别.

3. 举出一个你认为是统计应用的例子.

[1] statistics. (2008). Encyclopædia Britannica. Encyclopædia Britannica 2007 Ultimate Reference Suite. Chicago: Encyclopædia Britannica.

1.2 现实中的随机性, 概率和机会

从中学起, 大家就知道自然科学的许多定律, 例如物理中的牛顿三定律, 物质不灭定律以及化学中的各种定律等等. 但是在许多领域, 很难用如此确定的公式或论述来描述一些现象. 比如, 人的寿命是很难预先确定的. 一个吸烟、喝酒、不锻炼、而且经常吃荤的人可能比一个很少得病, 生活习惯良好的人活得长. 因此, 可以说, 活得长短有一定的**随机性(randomness)**. 这种随机性可能和人的经历、基因、习惯等无数不易说清的因素都有关系. 但是从总体来说, 我国公民的平均预期寿命却是非常稳定的, 而且随着生活水平的提高在逐步增长, 比如1996年的平均预期寿命为70.80岁, 而2000年为71.40岁. 这就是规律性. 一个人可能活过这个预期年龄, 也可能活不到这个年龄, 这是随机的. 但是总体来说, 预期寿命的稳定性, 却说明了随机之中有规律性. 这种规律就是统计规律.

你可能经常听到**概率(probability)**这个名词. 有一段时间在天气预报中常提到的降水概率. 大家都明白, 如果降水概率是百分之九十, 那就很可能下雨, 但如果是百分之十或者更少, 就不大可能下雨. 因此, 从某种意义说来, 概率描述了某件事情发生的机会. 显然, 这种概率不可能超过百分之百, 也不可能少于百分之零. 换言之, 概率是在0和1之间(也可能是0或1)的一个数, 说明某事件发生的机会有多大.

有些概率是无法精确推断的. 比如你对别人说你下一个周末去公园的概率是百分之八十. 但你无法精确说出为什么是百分之八十而不是百分之八十四或百分之七十八. 其实你想说的是你很可能去, 但又没有完全肯定. 实际上, 到了周末, 你或者去, 或者不去, 不可能有分身术把百分之八十的你放到公园, 而其余的放在别处. 有些概率是可以大体知道的. 比如掷骰子. 只要没有人在骰子上做手脚, 你得到6点的概率应该是六分之一. 得到其他点的概率也是一样. 这反映了掷骰子的规律性. 但掷出骰子之后所得到的结果还只可能是六个数目之一. 这体现了随机性. 如果你掷1000次骰子, 那么, 大约有六分之一的可能会得到6点, 这也说明随机结果也具有规律, 而且有可能通过试验等方法来推测其规律.

思考一下:

1. 有没有大于1或小于0的概率?

2. 举出若干可以计算的概率和无法计算的概率.

3. 有没有事物和概率无关?

1.3 变量和数据

做任何事情都要有对象. 比如一个班上注册的学生有200人, 这是一个固定的数目, 称为**常数(constant)**或者**常量**. 但是, 如果猜测今天这个班有多少人会来上课, 那就没准了. 这有随机性. 可能有请病假或事假的, 也可能有逃课

的. 这样, 就要来上课的人数是个**变量(variable)**. 另外对于某项政策同意与否的回答, 也有"同意"、"不同意"或者"不知道"三种可能值, 这也是变量, 只不过不是数量而已. 当变量按照随机规律所取的值是数量时该变量称为**定量变量**或**数量变量(quantitative variable)**, 因为是随机的, 也称为**随机变量(random variable)**. 像性别或观点之类的取非数量值的变量就称为**定性变量、属性变量**或**分类变量(qualitative variable, categorical variable)**. 这些定性变量也可以由定量变量来描述, 比如男性和女性的数目、同意某政策人数的比例等等. 定性变量只有用数量来描述时, 才能建立数学模型, 才能使用计算机来分析.

有了变量的概念, 什么是数据呢? 拿掷骰子来说, 掷骰子会得到什么值, 是个随机变量, 而每次取得1至6点中任意某点数的概率在理论上都是六分之一(如果骰子没有作假). 这依赖于在掷骰子背后的理论或假定, 而在实际掷骰子过程中, 如果掷100次, 会得到100个由1至6点组成的数字串, 再掷100次, 又得到一个数字串, 和前一次的结果多半不一样. 这些试验结果就是数据. 所以说数据是关于变量的观测值.

通过数据可以验证有关的理论或假定. 比如通过很多次掷骰子验证得到每个点的概率是不是1/6. 对于顾客是否喜欢某种饮品的调查也类似, 但这里不像掷骰子那样事先可以大致猜测顾客喜欢与否的概率. 在随机问了1000人之后, 可能有364人说喜欢, 而480人说不喜欢, 其余的人可能不回答, 或说不知道, 或从来没有喝过这种饮料. 当然, 它仅仅反映了1000个被问到的人的观点, 但这对于估计整个消费群体的观点还是有用的. 从这些数据可以估计出喜欢这种饮料的大约占364/1000=36.4%. 后面还要介绍得到数据的一些途径和方法.

思考一下:

1. 如果你抽签得到奖品的概率为十分之一, 但抽完签之后, 你或者得到奖品, 或者得不到, 这里就不存在概率了. 这实际上是一个随机实验的结果, 或者是一个随机变量(得到奖品与否)的实现值. 举例说明随机变量及其实现值之间的区别.

2. 数据是变量的实现值. 在概率论的文献中, 习惯上把随机变量用大写字母(比如X、Y)表示, 而把它们的实现值或数据用小写字母(比如x、y)表示.

1.4 变量之间的关系

现实世界的问题都是相互联系的. 不讨论变量之间的关系, 就无从谈起任何有深度的应用, 而没有应用, 统计的基本概念就仅仅是摆设而已.

人们每时每刻都在关心事物之间的关系. 比如, 职业种类和收入之间的关系、政府投入和经济增长之间的关系、广告投入和经济效益之间的关系、治疗手段和治愈率之间的关系等等. 这些都是二元的关系. 还有更加复杂的诸多变量之间的相互关系, 比如企业的固定资产、流动资产、预算分配、管理模式、生产率、债

务和利润等诸因素的关系是不能用简单的一些二元关系所描述的. 下面用例子探索性地说明变量之间可能存在的关系. 这些描述性的例子所涉及的统计方法都会在以后的章节中陆续介绍.

1.4.1 定量变量间的关系

例1.1 广告投入和销售之间的关系. 数据: ads.sav, ads.txt)显示了某企业的广告投入和销售额之间的关系(万元).

某企业广告投入(ads)和销售额(sales)数据(单位: 万元)

ads	1.00	3.20	3.20	5.50	5.90	7.10	7.30	9.20	10.80	12.10
sales	9.40	31.80	33.20	52.40	53.50	56.00	56.90	59.20	60.10	63.50

到底广告投入和销售额之间有没有关系? 还是用二维点图(称为散点图, 将在第三章介绍)来"感觉"一下这个数据. 图1.1的横坐标(ads)代表广告投入, 而纵坐标(sales)代表销售收入. 上面表格中的数字就由图中的点表示了. 从该图可以看出, 在广告投入少的时候, 广告投入和销售额之间有很强的相关, 广告投入增加, 销售额也增加, 但当广告投入达到一定水平之后, 销售额的增加就不那么快了.

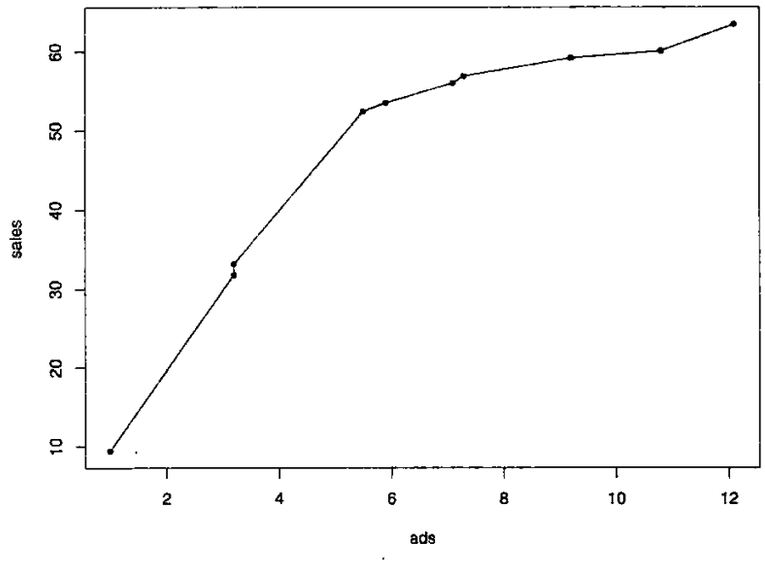

图 1.1 广告投入(ads)和销售额(sales)之间的关系.

一般来说, 人们希望能够从数据回答几个问题, 下面就此例进行初步探讨:

1. 这两个变量是否有关系? 看来, 它们有关系, 这从散点图就很容易看出. 看上去销售额是随着广告投入的递增而递增.
2. 如果有关系, 它们的关系是否显著? 这也可以从散点图得到. 当广告投入在6万元以下, 销售额增长很快, 但大于这个投入时, 销售额增长就不明显了. 因此, 这两个变量的关系是由强变弱.

3. 这些关系是什么关系? 是否可以用数学模型来描述? 本例看上去是可以拟合一个回归模型(后面会介绍), 但似乎不是线性的(用一条直线可以描述的). 具体细节需要进一步的分析.

4. 这个关系是否带有普遍性? 也就是说, 仅仅对这一个企业, 在这一段时间有这样的关系, 还是对于其他企业在其他时间也有类似的规律? 这里的数据还远远不足以回答这个问题. 可能需要考虑更多的变量和收集更多的数据. 一般来说, 人们希望能够从一些特殊的样本, 得到普遍的结论, 以利于预测.

5. 这个关系是不是因果关系? 这个问题可能永远不能准确地回答. 实际上, 销售额的增加很可能有多方面因素, 比如产品的改善, 销售渠道的畅通, 员工的管理, 成本的降低, 整个经济的改善, 购买力的提高等等, 说不定广告根本不起多大的作用, 这种关系仅仅是巧合而已. 严格来说, 只有排除了所有可能的影响因素之后, 才能讨论余下变量之间的因果关系. 而这种排除在实际生活中是几乎不可能的. 虽然如此, 在可控制的试验中, 特别是科学试验中, 还是有可能找到必要的因果关系的. 但是, 一般来说, 变量之间有关系这个事实并不意味着一定存在明确的因果关系. 比如肺癌和吸烟肯定是相关的, 但是, 有人认为由于某种不明原因或其他一些变量造成了这二者同时出现, 至少, 吸烟并不是得肺癌的充分条件. 再例如, 任何和时间有关的变量, 都有可能有某种相关, 比如一个婴儿的体重增长, 和同时期的国民经济的增长就很可能相关, 但没有人会认真建立婴儿的增长和经济增长的模型(当然不妨试试). 然而, 只要有关系, 即使不是因果关系也不妨碍人们利用这种关系来进行推断. 也有人认为, 较早发生的事件为原因, 而后发生的为结果, 但公鸡打鸣在先, 太阳升起在后、地震先兆在前, 地震在后, 这都不能说明发生的时间先后能够成为判断因果关系的依据.

上面列出的这些问题并不是一成不变的, 也不是每个问题都需要回答或者能够得到答案的. 一切根据实际需要和手中掌握的数据而定. 简单的办法(诸如上面的散点图)往往不一定能够给出满意的答案, 这就需要更多的工具和手段来进行数值分析, 以得到更加严格和精确的解答.

这里可能有必要说明, 日常用语"关系"一词是没有严格统计定义的, 统计术语"相关" (当然也是日常用语)试图用统计语言来描述一些关系. 但目前的统计"相关"仅仅描述了日常所说的"关系"的很小的一部分. 这就好像宇宙是无穷的, 而人类的科学理论或假说只能覆盖很小的部分一样. 正如上面所提到的, 有些关系是直接的因果关系, 比如增加热量可以加速金属的熔化、加催化剂可以加速某些化学反应. 有些关系则看不出哪个是因, 哪个是果, 或者都是某个共同原因的结果, 比如, 高血压和动脉硬化是相关的, 但它们很可能都是整个机体的某种状态的表现, 是基因、环境等许多因素的结果. 中医和西医对疾病处理的不同就反映了他们对因果的不同看法. 再例如, 一个母亲的婴儿生长和另一个母亲的婴儿生长显然是有关系的, 但这不是直接的因果关系, 这是人类的共同基因决定的. 虽然没有直接关系, 但一个母亲很容易看出其他婴儿的年龄, 这也是一种基于自己婴儿的模型对其他婴儿进行的一种推断. 另外一些关系看上去可能没有一个共同的原因,

但可能有类似的规律, 比如随着时间进程而出现的植物和动物的生长、经济的增长、人口的增长等等. 当然, 即使不用人口增长来预测经济增长, 人们也很难绝对地说这些事物真正没有共同的影响因素, 比如, 环境的破坏会把上述生长或增长全部破坏. 当然, 必定存在一些纯粹的巧合, 但这些偶然相关是不会形成规律的, 如果有了规律, 那就不是巧合了.

思考一下:

1. 少数数据所展示的关系很可能是偶然的. 必须区分偶然事件和有规律事件之间的区别, 请举例说明你对这个问题的理解.
2. 是不是两个事件频繁地、固定地先后发生的现象可以证明: 先发生的事件为后发生事件的原因? 举例说明.

1.4.2 定性变量间的关系

例1.2 (数据: change.txt) 这是对某地区一个行业员工的调查数据中三个问题所组成的列联表. 这里的三个问题是: "你的年龄"(三个范围选一项: 在数据中代码(哑元)1代表小于30岁, 2代表30-40岁, 3代表40岁以上), "你的教育程度"(三个范围选一项: 在数据中代码1代表"本科及以上", 2代表"专科", 3代表"专科以下"), "你是否想跳槽"(三个范围选一项: 在数据中代码1代表"想跳槽", 2代表"不想跳槽", 3代表"不知道"). 下表是涉及这三个问题的列联表. 注意, 实际计算机软件从文件中所读入的数据形式和此表不大相同, 这个表是计算机转换出来的. 一般原始数据的形式为常用的方阵形式, 如本数据在change.txt中的形式.

员工调查数据

	是否想跳槽(Change)	想跳槽(1)			不想跳槽(2)			不知道(3)		
	教育(Edu)	1	2	3	1	2	3	1	2	3
年	<30岁(1)	28	110	70	4	12	19	10	59	64
	30-40(2)	31	138	67	11	27	18	20	99	89
龄	>40(3)	2	14	23	1	4	11	5	18	36

注: 教育的哑元中, 1代表至少本科, 2代表专科, 3代表专科以下.

这种数据每个变量都有几种取值, 比如年龄有三个可能取的值(30岁以下, 30-40岁和40岁以上), 称为三个水平(level). 类似地, 教育有三个水平, 是否想跳槽有三个水平,这个表中间的数目是被调查人相应于变量各种水平组合(共有$3\times3\times3=27$种组合)出现的频数. 比如, 小于30岁本科及以上想跳槽的为28人. 大于40岁专科以下想跳槽的为23人等等. 从这个表中, 还可以算出一些部分和. 比如想跳槽的有483人, 总人数999人, 本科及以上教育程度的人有112人等等. 这个表不如前面例1.1的散点图那么直观. 下表为仅仅保留教育程度和是否想跳槽这两个变量的列联表:

仅有两个变量的员工调查数据列联表

	是否想跳槽	跳槽	不跳槽	不知道
教	大学	61	16	35
	大专	262	43	176
育	中小	160	48	189

从这个列联表能够看出学历高的比较学历低的更想跳槽吗?

思考一下:

1. 定性变量之间的关系也要由数量表示,这就是不同变量不同水平的组合中事件发生的数目或频数. 能不能完全不用数目来描述定性变量之间的关系? 请举例说明.

2. 完全是定性变量的数据倒可以仅仅由字符表示(不一定是哑元),如果你知道,请举例.

1.4.3 定性和定量变量间的混和关系

例1.1和例1.2中的变量类型比较单一. 下面看不同类型变量混和的例子.

例1.3 出生婴儿数据(lowbwt.txt). 该数据摘自Hosmer and Lemeshow (2000)[1], 给出了189个初生婴儿的重量(BWT), 母亲的年龄(AGE), 怀孕前母亲的重量(LWT), 母亲小产次数(PTL, 0, 1,... 等整数), 母亲头三个月中就医次数(FTV, 0, 1,... 等整数)等定量变量, 以及婴儿是否过重(LOW, 哑元0表示大于等于2500克, 1代表小于2500克), 母亲是否吸烟(SMOKE, 哑元1代表"是", 0代表"否"), 母亲是否有高血压(HT, 1代表"有", 0代表"没有"), 母亲是否有子宫过敏(UI, 1代表"有", 0代表"没有"), 及识别号码(ID)等定性变量. 人们试图从这个数据找出婴儿重量和各种定性和定量变量之间的关系. 注意, 这个数据中的婴儿是否过重(LOW)是从婴儿重量得到的, 不能用后者来推断前者.

1.5 统计、计算机与统计软件

现代生活越来越离不开计算机了. 最早使用计算机的统计当然更离不开计算机了. 事实上, 最初的计算机仅仅是为科学计算而设计和建造的. 大型计算机的最早一批用户就包含统计. 现在, 统计仍然是进行数字计算最多的用户之一. 当然计算机现在早已脱离了仅有数字计算功能的单一模式, 而成为百姓生活的一部分. 计算机的使用, 也从过去必须学会计算机语言到只需要"傻瓜式"地点击鼠标. 结果也从单纯的数字输出到包括漂亮的表格和图形在内的各种形式.

统计软件的发展, 也使得统计从统计学家的圈内游戏变成了大众的游戏. 只要

[1]Hosmer and Lemeshow (2000) Applied Logistic Regression: Second Edition. These data are copyrighted by John Wiley & Sons Inc. and must be acknowledged and used accordingly. Data were collected at Baystate Medical Center, Springfield, Massachusetts during 1986.

输入你的数据, 点几下鼠标, 做一些选项, 马上就得到令人惊叹的漂亮结果了. 人们可能会问, 是否傻瓜式统计软件的使用可以代替统计课程了? 当然不是. 数据的整理和识别, 方法的选用, 计算机输出结果的理解都不像使用傻瓜相机那样简单可靠. 有些诸如法律和医学方面的软件都有不少警告, 不时提醒你去咨询专家. 但统计软件则不那么负责. 只要数据格式无误、选项不矛盾而且不用零作为除数就一定给你结果, 而且几乎没有任何警告. 另外, 统计软件输出的结果太多, 即使是同样的方法, 不同软件输出的内容还不一样, 有时同样的内容名称也不一样. 这就使得使用者大伤脑筋. 即使是统计学家也不一定能解释所有的输出. 因此, 就应该特别留神, 明白自己是在干什么. 不要在得到一堆毫无意义的垃圾后还沾沾自喜.

统计软件的种类很多. 本书采用免费的自由编程软件R来实现我们的目标, 读者可以毫不费力地重复书中所有的例题的计算. R软件从1995年问世以来, 已经成为世界统计学家的首选研究和教学软件. R网站[1] 拥有世界各地统计学家贡献的大量最新程序包(package), 这些程序包以飞快的速度增加和更新, 已从2009年底的大约1000个增加到2012年8月底的4009个, 仅2012年8月份就增加了449个. 它们代表了统计学家创造的崭新的统计方法. 这些程序包的代码都是公开的[2]. 与此相对比, 所有商业软件远没有如此多的资源, 也不会更新得如此之快, 而且商业软件的代码都是保密的昂贵"黑匣子". 在发达国家, 不能想象一个统计研究生不会使用R软件. 那里很多学校都开设了R软件的课程. 今天, 任何一个统计学家想要介绍和推广其创造的统计方法, 都必须提供相应的计算程序, 而发表该程序的最佳地点就是R网站. 由于方法和代码是公开的, 这些方法很容易引起有关学者的关注, 这些关注对研究相应方法形成群体效应, 推动其发展. 不会编程的统计学家在今天是很难生存的.

在学校中讲授任何一种商业软件都是为该公司做义务广告, 如果没有相关软件公司的资助, 就没有学校愿意花钱讲授商业软件. 在教学中使用盗版软件是违法行为, 绝对不应该或明或暗地鼓励师生使用盗版商业软件.

无论从编程逻辑还是技巧上, 对R软件编程的熟悉无疑有助于学习其他快速计算的编程语言, 比如C++和FORTRAN, 这对于应对因快速处理庞大的数据集而面临的巨大的计算量有所裨益.

R软件安装和运行小贴士

- 登录R网站(`http://www.r-project.org/`)[3], 根据说明从你所选择的镜像网站来下载并安装R的所有基本元素.
- 向左边变元赋值语句可以用"="号或者"<-"; 还可以用"->"向右赋值.
- 运行时可以在提示码">"后逐行输入指令. 如果回车之后出现"+"号, 则说明你的语句不完整(得在+号后面继续输入)或者已输入的语句有错误.

[1] 网址: http://www.r-project.org/.
[2] 除了极个别并非秘密的子程序之外. 因为它们很费时间, 用机器代码实行.
[3] 网上搜索"R"即可立得到网址.

- 每一行可以输入多个语句, 之间用半角分号";"分隔.
- 所有代码中的标点符号都用半角格式(基本ASCII码). R的代码对于字母的大小写敏感. 变量名字、定性变量的水平以及外部文件路径和名字都可以用中文.
- 不一定非得键入你的程序, 可以粘贴, 也可以打开或新建以R为扩展名的文件(或其他文本文件)作为运行脚本, 在脚本中可以用Ctrl+R 来执行(计算)光标所在行的命令, 或者仅运行光标选中的任何部分.
- 出现的图形可以用Ctrl+W或Ctrl+C来复制并粘贴(前者像素高), 或者通过菜单存成所需的文件格式.
- 如果在运行时点击Esc则会终止运行.
- 在运行完毕时会被问到"是否保存工作空间映像?", 保存的结果是下次运行时, 这次的运行的结果还会重新载入内存, 不用重复计算, 缺点是占用空间. 如果已经有脚本, 而且运算量不大, 一般都不保存. 如果你点击了保存, 又没有输入文件名, 这些结果会放在所设或默认的工作目录下的名为.RData的文件中, 你可以随时找到并删除它.
- 注意, 从ppt或word文档等各种非文本文件中复制并粘贴到R上的代码, 则可能存在由这些软件自动变换的首字大写或者左右引号等造成的R无法执行的问题.
- R中有很多常用的数学函数、统计函数以及其他函数. 可通过在R的帮助菜单中选择"手册(PDF文件)", 在其附录中找到各种常用函数的内容.
- 在R界面, 你可以用问号加函数名(或数据名)来得到该函数或数据的细节, 比如用"?lm"可以得到关于线性模型函数"lm"的各种细节. 另外, 如果想查看在MASS程序包中的稳健线性模型"rlm", 在已经打开该程序包时(用library(MASS)打开, 用detach(package:MASS)关闭), 可用"?rlm"来得到该函数的细节. 如果MASS没有打开[1], 或者不知道rlm在哪个程序包, 可以用"??rlm"来得到其位置. 如果对于名字不清楚, 但知道有部分字符, 比如"lm", 可以用"apropos("lm")"来得到所有包含"lm"字符的函数或数据.
- 如果想知道某个程序包有哪些函数或数据, 则可以在R的帮助菜单上选择"Html帮助", 再选择"Packages"即可得到你的R上装载的所有程序包. 这个"Html帮助"很方便, 可以链接到许多帮助(包括手册等).
- 有一些简化的函数, 如加减乘除乘方("+, -, *, /, ^")等, 可以用诸如"?"+""这样的命令得到帮助(不能用"?+").

[1]通常为了节省内存以及避免变量名字混杂, 应该在需要时打开相应的程序包, 不需要时关闭.

- 你还可以写关于代码的注释: 任何在"#"号后面作为注释的代码或文字都不会参与运行.

- 你可能会遇到无法运行过去已经成功运行过的一些代码, 或者得到不同结果的现象. 原因往往是这些程序包经过更新, 一些函数选项(甚至函数名称和代码)都已经改变, 这说明R软件的更新和成长是很快的. 解决的办法是查看该函数, 或者查看提供有关函数的程序包来探索一下究竟.

- 有一个名为RStudio的自由下载软件可以更方便的用几个窗口来展示R的执行、运行历史、脚本文件、数据细节等过程.

1.6 小结

这一章主要描述了统计领域的轮廓, 还说明了随机性所可能包含的规律性. 概率是对不确定性的度量. 统计研究的对象是变量. 有了变量, 特别是随机变量, 才能够有目的地收集与该变量有关的数据, 对数据进行分析, 并且得到人们感兴趣的结论. 单独变量的研究很重要, 但应用中人们最关心的是变量之间的关系. 研究各种变量之间的关系占了本书的大部分内容. 为了进行数量分析, 使用计算机是不可避免的. 现代应用统计是离不开计算机的. 对于非统计工作者来说, 能够使用顺手的统计软件来处理数据是非常重要的. 有许多统计软件可供选择. 同时还要清醒地认识到, 如果选择了错误的方法或选用了无关的变量, 就不可能得到有用的结论. 计算机可以是人们的助手, 但不能代替人们的思维.

1.7 习题

1. 举出你所知道的统计应用例子.

2. 举出日常生活中随机性和规律性的例子.

3. 掷一个骰子, 或者抛一个钱币100次, 记录下结果, 并用此来解释随机性和规律性以及概率的概念.

4. 你使用过统计软件或者利用过其软件中的统计功能吗? 你有什么经验和体会?

5. 举出有若干定量变量的(假想的或真实的)例子. 说出你希望得到的结论.

6. 举出有若干定性变量的(假想的或真实的)例子. 说出你希望得到的结论.

7. 举出既有定性变量又有定量变量的(假想的或真实的)例子. 说出你希望得到的结论.

8. 举出任何涉及变量关系的例子.

9. 举例讨论各种变量的因果关系.

10. 搜寻到R网站(http://www.r-project.org/), 并在CRAN (The Comprehensive R Archive Network)的镜像网站(CRAN mirror)下载R的基本软件(Base). 在你的计算机上完全安装R软件. 然后阅读"帮助"中的手册(PDF文件). 你就可以试着一步一步地自学R软件的使用了.

11. 打开R软件, 通过选项文件⇒打开程序脚本找到光盘中的文件"R练习.R"[1], 再用其中提供的语句尝试R软件, 每执行一两行, 观察输出, 再思考一下. 自己体会其中的规律. 好, 慢慢品尝R的奥妙吧! (请参看上面"R软件安装和运行小贴士").

[1]这个练习也附在本书后面.

第二章 数据的收集

统计的对象是世界上的各种问题, 要得到人们感兴趣的问题所包含的规律, 就必须收集与问题相关的信息, 也就是数据. 因此, 在收集数据之前, 必须根据问题的性质, 找到相关的变量, 然后再收集这些变量的观测值. 寻找相关变量所需要的是相应领域的知识, 统计知识本身是不够的. 只有对相关变量的数据做出分析, 才能得到有价值的结论.

2.1 数据是怎样得到的?

翻开报纸、打开电视或网页, 就可能看到各种数据. 比如就业率、高速公路通车里程、物价指数、股票行情、外汇牌价、犯罪率、房价、流行病等有关数据, 还有包括统计局系统及各个政府机构定期发布的各种国家经济数据、进出口贸易数据及税务等等. 从这些数据中, 各有关方面可以提取对自己有用的信息. 这些间接得到的数据都称为二手数据.

获得第一手数据并不像得到二手数据那么轻松. 某些企业每年至少要花三四千万元来收集和分析数据. 他们调查其产品目前在市场中的状况和地位并确定其竞争对手的态势; 他们调查不同地区、不同阶层的民众对其产品的认知程度和购买意愿, 以改进产品或推出新品种以争取新顾客; 他们还收集各地方的经济交通等信息, 以决定如何保住现有市场和开发新市场. 市场信息数据对企业是至关重要的. 他们很舍得在这方面花钱. 因为这是企业生存所必需的, 不能是可有可无.

上面所说的数据是在自然的未被控制的条件下观测到的, 称为**观测数据(observational data)**. 而对于有些问题, 比如在不同的医疗手段下某疾病的治疗结果有什么不同, 在不同的肥料和土壤条件下某农作物的产量有没有区别, 用什么成分可以提高某超导材料的温度等等. 这种在人工干预和操作情况下收集的数据就称为**试验数据(experimental data)**.

思考一下:

1. 试图想象你自己如何收集关于周围人群购买习惯的数据. 需要什么变量?

2. 在媒体上出现的数目中多为单个数目, 从单个数目能够得到规律吗?

2.2 个体、总体和样本

要想了解北京市民对建设北京交通设施是以包括轨道运输在内的公共交通工具为主还是以小汽车为主的观点, 需要进行调查, 调查对象是所有北京市民, 调查目的是希望知道市民中对这个问题的不同看法各自占有的比例. 显然, 不可能去问所有的北京市民, 而只能够问一部分, 并且根据这一部分的观点来理解整个北京市民的总体观点. 在这个例子中, 单个北京市民称为调查的**对象(object)**, 而他们的观点称为(这个调查问题中)的**个体(element, individual, unit)**, 而称

所有北京市民对这个问题的观点为一个**总体(population) 或有限总体(finite population)**[1], 总体是包含所有要研究的个体的集合. 而调查时问到的那部分市民的观点(也就是部分个体)称为该总体的一个**样本(sample)**, 是总体中选出的一部分. 当然, 也有可能试图调查所有的人(比如人口普查), 那叫做**普查(census)**. 有人喜欢把作为调查对象的北京市民称为个体, 但一个市民还有其他诸如身高、体重、收入、职业、教育程度等大量其他特征, 这些都不是这个调查的目的. 实际上, 市民本身是调查对象, 而市民的观点才应称为个体.

在抽取样本时, 如果总体中的每一个体都有同等机会被选到样本中, 这种抽样称为**简单随机抽样(simple random sampling)**, 而这样得到的样本则称为**随机样本(random sample)**. 就北京交通问题的调查为例, 在简单随机抽样的情况下, 如果**样本量(sample size)**, 也就是样本中个体的数目在总体中的比例为1/5000, 那么, 无论在东城区或者在延庆县, 无论在白领阶层还是蓝领阶层被问到的人的比例都应该大体是1/5000. 也就是说, 这种比例在总体的任何部分是大体不变的. 换言之, 在随机抽样的一个样本中各个不同特征人群的比例和他们在总体中的比例应该类似. 随机抽样这就像从一锅搅和均匀的八宝粥中舀出一勺, 其中各种成分的比例应该和锅里的比例大致一样.

大小为N的总体中产生样本量为n的随机样本的一个常用的方法是利用**随机数(random number)** [2], 其步骤为: (1)先把总体的所有个体编号, (2)然后产生n个在0到N之间的随机数, (3)与如此产生的随机数中的数目相同的个体则形成了样本量为n的简单随机样本. 那么, 如何获得随机数呢? 最原始的办法是掷一种正20面体的均匀材料制成的骰子, 其20个表面标有两套0到9的数字. 每掷一次产生一个0到9的数字. 假定总体大小$N = 1200$, 而样本量$n = 50$. 人们可以掷这个骰子4次(或者掷4个不同颜色的骰子, 每个颜色代表一位数)产生一个4位数目. 这样不断掷下去直到得到50个在1和1200之间的数目. 这就是所需要的随机数. 另一种得到随机数的方法是查阅随机数表. 在一些传统的统计教科书中可以找到随机数表, 也有专门的随机数表的册子. 随机数表的数目无论从页数、行或列来看都是随机的. 比如$N = 1200$, 而$n = 50$, 那么, 在随机数表中可以取4列, 然后往下找到50个在1和1200之间的数目即可. 当然, 用随机数表产生随机数的方式还有很多. 这里不多讲. 在广泛使用计算机的今天, 为了方便, 很多实际工作者应用计算机所产生的**伪随机数(pseudo-random number)**来代替真正的随机数. [3]

在实践中, 得到随机样本不容易. 很多搞调查的人就采取简单的办法. 还以北京的交通问题的调查为例. 如果按照随机选出的电话号码进行调查, 则肯定节省时间和资源, 但这样得到的就不是一个随机样本了. 首先, 没有电话的阶层就不会被问到. 另外, 如果号码是从住户号码中选, 那么白天打住户电话, 得到的多半是

---

[1]注意, 这里的术语总体和与概率分布结合的总体(样本空间)概念有所不同. 但如果确定了抽样方法, 那么这里的总体中所感兴趣的个体的个数或比例则可以用后面的总体概念中的分布参数来描述. 请参看本章总结一节中的注.

[2]这里所说的随机数的意义是指任何在n个在0到N之间的数目都有同等的机会被选中. 更广义的随机数是独立同分布的随机变量的实现. 参见后面第四章.

[3]用R软件可以产生各种分布的随机数, 比如用x=runif(100, 2,3)可以产生100个2和3之间的伪随机数.

白天不在单位工作的人的意见. 即使都在家, 那一家人无论多少口人一般就只有接电话人的观点被调查到. 这一类的样本称为**方便样本(convenience sample)**. 在调查中, 即使选择对象的确是随机的, 最理想的情况所得到的样本也只代表那些愿意回答问题人的观点所组成的总体, 而不愿回答问题的人的观点永远不会得到. 这种不回答所造成的问题是抽样调查特有的问题. 在其他问题中, 也有使用方便样本的情况. 比如在肺癌研究中, 人们往往看到吸烟和肺癌关系的数据, 这些数据多半不是从整个人群中采集的随机样本, 它们可能只是医院中的病人记录中得到的. 在杂志和报纸上也有问卷, 但得到的只是拥有这份报刊, 而且愿意回答的人的观点.

思考一下:

1. 在街上向随机遇到的人提问, 这样得到的样本是随机样本吗? 在什么限定条件下它可能是随机样本?

2. 网上有许多调查, 这些调查关于什么总体可以说是随机样本?

2.3 收集数据时的误差

假定在某一职业人群中女性占的比例为60%. 如果在这个人群中抽取一些随机样本, 这些随机样本中女性的比例并不一定刚好是60%, 可能稍微多些或稍微少些. 这是很正常的, 因为样本的特征不一定和总体完全一样. 这种差异不是错误, 而是必然会出现的抽样误差(sampling error). 刚才提到在抽样调查中, 一些人因为种种原因没有对调查做出反应(或回答), 这种误差称为**未响应误差(nonresponse error)**. 而另有一些人因为各种原因回答时并没有真实反映他们的观点, 这称为**响应误差(response error)**. 和抽样误差不一样, 未响应误差和响应误差都会影响对真实世界的了解, 应该在设计调查方案时尽量避免.

2.4 抽样调查和一些常用的方法

抽样调查(sampling survey)的领域涉及如何用有效的方式得到样本数据. 最常用的问卷调查方式包括通过邮件报刊网络等手段调查、电话调查和面对面调查等. 这些调查都利用了**问卷(questionnaire)**, 而问卷的设计则很有学问. 它涉及如何用词、问题的次序和问题的选择与组合等等. 这涉及包括心理学、社会学等知识. 面对面调查则需要对调查者进行培训. 首先, 问卷中的问题数目不能太多. 太多了, 回答者就会厌倦, 而不能得到真实结果. 为了提高效率, 问题一般都是选择题, 但选择项不宜过多. 问题的语言应该和被调查者的文化水平相适应, 通俗易懂, 但又要准确而不至于造成误解. 笔者曾经见到一个失败的问卷, 后来在研究生课上让他们理解该问卷的问题, 结果多数研究生不能理解或者做相反的理解. 有时本来被访者没有观点, 但问卷的措辞使得被访者觉得一定要选择一个观点. 问题的次序也很重要, 简单的在先, 等到"热身"以后, 再提敏感的和核心的问题, 这在面对面调查时尤为重要. 另外, 注意问题的相关性可能会使人觉得必须前

后一致, 比如, 在前面问题中, 被访者回答说支持公共交通, 而在后面问是否购买小轿车时, 可能就会犹豫, 觉得应该回答"不买"才和前面一致, 其实这两个问题并不必要联系起来. 在面对面调查中, 调查者的选择也很重要, 不能想象, 一个西装革履的调查者能够从贫困人群中对某些敏感问题得到真实可信的回答. 也有不包含问卷的抽样调查, 比如, 对个人或企业的信用记录的抽样, 对个人或企业的纳税记录的抽样等, 也可以用计算机从大型数据库来抽样.

抽样调查的设计的目的之一是确保样本对总体的代表性, 以保证后继推断的可靠性. 前面说到, 每个个体等可能的简单随机抽样是一个理想情况. 这种简单随机抽样是**概率抽样方法(probability sampling method)**的一个特例. 概率抽样假定每个个体出现在样本中的概率是已知的. 这种概率抽样方法使得数据能够进行合理的统计推断. 但是为了节省调查的费用和时间, 常常采取基于方便或常识判断的**非概率抽样方法(nonprobability sampling method)**. 对从非概率抽样得到的数据进行推断要非常慎重. 它依赖于具体的抽样方案是如何设计的, 也依赖于它是如何实施的. 这种推断往往无法根据漂亮的统计理论来进行. 也很难客观地建立抽样误差的范围.

在抽样调查时, 最理想的样本是前面提到的简单随机样本. 但是由于实践起来不方便, 在大规模调查时一般不用这种全部随机抽样的方式, 而只是在局部采用随机抽样的方法. 下面介绍几种抽样方法. 这里没有深奥的理论, 读者完全可以根据常识判断在什么情况下无法获取简单的随机样本, 以及下面的每个方法有什么好处和缺陷. 对于它们具体的设计、实施与数据分析, 有许多专门的书籍, 就不在这里赘述了. 另外, 一般仅有少数人有机会来确定抽样方案. 读者仅需把这些方法当成常识来了解就可以了. 下面是一些概率抽样方法.

1. **系统抽样(systematic sampling)**. 这也称为每n个名字选择方法(n-th name selection technique)这是先把总体中的每个单元编号, 然后随机选取其中之一作为抽样的开始点进行抽样. 根据预定的样本量决定"距离"n. 在选取开始点之后, 通常从开始点开始按照编号进行所谓等距抽样. 也就是说, 如果开始点为5号, "距离"为$n = 10$, 则下面的调查对象为15号、25号等等. 不难想象, 如果编号是随机选取的, 则这和简单随机抽样是等价的.

2. **分层抽样(stratified sampling)**. 这是简单随机抽样的一个变种, 先把要研究的总体按照分成**相对相似或相对齐次(relatively homogeneous)** 的个体组成的**类(stratum)**, 再在各类中分别抽取简单随机样本. 然后把从各类得到的结果汇总, 并对总体进行推断. 在每类中调查的人数通常是按照该类人的比例, 但出于各种考虑, 也可能不按照比例, 也可能需要加权(加权就是在求若干项的和时, 对各项乘以不同的系数, 这些系数的和通常为1). 比如在一项教育程度可能和某些结果有关的调查中, 可以按照教育程度把要访问的人群分成几类, 再在每一类中调查和该类成比例数目的人. 这样就确保了每一类都有相应比例的代表. 分层抽样的一个副产品就是同时可以得到各类的结果.

3. **整群抽样(cluster sampling)**. 这是先把总体划分成若干**群(cluster)**. 和分

层抽样不同, 这里的群是由不相似或**异类的(heterogeneous)**个体组成. 在**单级整群抽样(single-stage cluster sampling)**中, 先(通常是随机地)从这些群中抽取几群, 然后再在这些抽取的群中对个体进行全面调查. 在**两级整群抽样(two-stage cluster sampling)**中, 先(通常是随机地)从这些群中抽取几群, 然后再在这些抽取的群中对个体做简单随机抽样. 比如, 在某县进行调查, 首先在所有村中选取若干村子, 然后只对这些选中的村子的人进行全面或抽样调查. 显然, 如果各村情况差异不大, 这种抽样还是方便的. 否则就会增大误差了. 整群抽样的主要应用是所谓**区域抽样(area sampling)**, 那时, 群就是县、镇、街区或者其他适当的关于人群的地理划分.

4. **多级抽样(multistage sampling)**. 在群体很大时, 往往在抽取若干群之后, 再在其中抽取若干子群, 甚至再在子群中抽取子群, 等等. 最后只对最后选定的最下面一级进行调查. 比如在全国调查时, 先抽取省, 再抽取市地, 再抽取县区, 再抽取乡、村直到户. 在多级抽样中的每一级都可能采取各种抽样方法. 因此, 整个抽样计划可能比较复杂, 也称为多级混和型抽样.

非概率抽样的方法的例子有:

1. **目的抽样(purposive sample)**. 这是由研究人员主观地选择对象. 比如在民意调查中, 在富人、中产阶级、穷人的街区各取得一些样本, 样本多少依赖于预先就有的知识.

2. **方便抽样(convenience sampling)**. 它用于探索性的研究, 研究人员以较少的花费得到对客观情况的近似. 这种非概率抽样常用于初期的评估. 比如, 你为了调查游客的意见, 你可以选择不同的时间和旅游景点, 随意对愿意停下的游客进行调查. 有时看起来是随机, 但实际上不是.

3. **判断抽样(judgment sampling)**. 研究人员凭经验判断来选择样本, 它通常是方便抽样的延伸. 比如要研究各县的情况, 而研究人员仅在一个县中抽样, 认为该县能够代表其他县.

4. **定额抽样(quota sampling)**与概率抽样中的分层抽样类似. 先是确定各类及比例, 然后利用方便抽样或判断抽样来从每一类中按比例选取需要的个体数.

5. **雪球抽样(snowball sampling)**. 它用于感兴趣的样本特征较稀有的情况, 依赖于一个目标推荐另一个目标的方法, 比如想要调查吸毒者的情况, 你先找到一个和吸毒者有关的人, 然后他(她)会介绍你找到其他的人. 虽然减少了花费, 但可能产生较大偏差.

6. **自我选择(self-selection)**. 这是让个体自愿参加调查. 比如对高血压病防治的调查, 一些人会作为自愿者来参加.

实际上的抽样通常都可能是各种抽样方法的组合. 既要考虑精确度, 还要根据客观情况考虑方便性、可行性和经济性. 不能一概而论.

思考一下:

1. 在一个大学中按照学号随机抽取一些学生来调查是不是随机样本, 在什么情况下是?
2. 按照男女比例, 在男女生中随机抽样属于什么类型的抽样?
3. 在学校中随机抽取一些系, 再随机抽取一些班级, 然后再进行随机抽样, 这是哪一类抽样?

2.5　计算机中常用的数据形式

数据是由一些变量和它们的观测值所组成. 在第一章的例子已经介绍了一些数据. 下表的例子就是调查人们对某个问题观点的一个数据的方阵形式. 其中有6个变量: 观点(观测值为支持、反对和不知道三种)、教育程度(有高中低三种取值, 用H、M、L表示)、月收入(取值为实际数字)、性别(取值有男女两个, 用M和F表示)以及地区号(用数字1, 2, 3, 4表示)等. 该表一共有1364个观测值(问卷回答). 可以看出这些变量有定性(属性)变量, 也有定量(数值)变量. 按照这个数据的格式, 每一列为一个变量的不同观测值, 而每一行则称为一个观测值, 它是个由数量值和属性值组成的向量, 每一个值相应于一个变量.

对某项政策的观点调查的原始数据形式

被访者编号	观点	教育程度	月收入	性别	地区号
1	支持	H	1600	M	1
2	支持	M	1720	F	1
3	反对	L	700	M	1
4	支持	H	2000	F	2
5	不知道	M	1000	M	2
6	不知道	L	600	F	1
⋮	⋮	⋮	⋮	⋮	⋮
1363	反对	L	1080	M	4
1364	支持	H	2100	M	3

还可以抽取该数据的部分形成一些汇总表格, 供在文献中研究和展示之用, 但汇总后表格(如上表)都不是计算机常用的数据形式. 这些汇总表格简单明了, 是通常媒体上最常见的形式之一. 但是从高维表汇总来的低维表不能还原成原始的高维表, 肯定损失一些有用的信息. 因此, 在做分析时, 尽量使用原始数据, 汇总加工过的数据信息损失很大, 一般只能作为最终展示结果, 而不宜作为原始材料来做数据分析.

对于比较复杂的问题, 一个数据可能由多个数据文件组成, 或者有特别的格式, 不一定都是单一的方阵形式.

思考一下：

1. 最彻底的汇总是把每个变量的所有数据求和，或者求平均，得到一个数目，但这就没有任何做进一步统计分析的余地了.

2. 统计部门公布的许多数据都是汇总数据，仅仅是为展示而用. 请分别举出一些原始数据和加工过的数据的例子. 并说明在加工过的数据中，哪些信息永远失去了.

2.6 小结

本章概述了数据的获取. 数据可以按照是否可以控制有关因素划分为试验数据和观测数据，也可以根据数据获得的途径划分为一手数据和二手数据. 数据往往是从总体中抽取出来的，所以，它是总体的一个代表，称为样本. 样本有简单随机样本，也有方便样本，依抽样时采取的方法而定. 收集数据时，会有必然出现的不是错误的抽样误差，也可能出现在调查实践中应该避免的未响应误差和响应误差. 本章还介绍了抽样调查和在抽样调查时常用的几种抽样方法，即分层抽样、整群抽样、多级抽样和系统抽样等. 实际的抽样过程可能是这些抽样的组合. 一句话，收集数据或抽取样本是为了从样本中得到总体的信息. 因此数据收集是否妥当，关系到后继分析和推断的结果是否合理. 最后，还介绍了常用的计算机使用的数据形式. 主要形式是由变量和观测值组成的方阵形式.

关于总体术语的不同概念的注

后面有关于总体参数的推断的内容，那里的总体(样本空间)和抽样调查中的原始的不包含概率分布(也就是不涉及抽样方法)的有限总体有所不同. 为了不使读者困惑，这里加上关于总体概念的注，并把这里的有限总体根据不同的抽样实践和超几何分布或二项分布结合，使得对抽样调查涉及的有限总体中感兴趣个体的个数或比例可以通过对后面意义上的总体参数的推断来研究. 这里涉及的一些概率分布可在第四章找到.

注意，前面在抽样调查例子中引进的术语“总体”是人们所关心的所有个体的集合(成为有穷的样本空间)，也称为有限总体. 这个总体是一个客观存在的事物，和人们抽样的实践无关. 此外，术语总体还在概率论中被用来表示另外一个概念，这就是**样本空间(sampling space)**，它定义为所有和某个试验相关的基本事件的集合. 任何不可分解的试验结果被一个而且仅仅被一个样本空间的点(称为**样本点，sample point**)所代表. 样本空间是一个抽象的集合，包括了定义在其子集的$\sigma-$代数上的概率测度[1]. 比如，用通俗的话来说，一个正态总体包括了一个服从某正态分布的试验(比如对某个物体用精密天平称重)的所有可能取值范围及该正态分布本身. 而一个由n个Bernoulli试验定义的二项总体则包括了这些试验的所有可能取值(在n次试验中可能成功的次数，即0到n之间的整数)加上该二项分布本

[1]这一句中的$\sigma-$代数和概率测度为高等概率论的术语，读者不必在意，写在这里仅仅是为了让教师和感兴趣的读者感到概念的完整性.

身. 这都是理论上的总体, 用来描述随机现象的规律性. 这些总体都有总体参数, 如正态分布中的均值μ和方差σ^2, 二项分布中Bernoulli试验的成功概率p等等. 这些总体参数可以通过试验结果(样本)的一些统计量来进行推断.

如果确定了抽样方法, 则前面提到的抽样调查中的(有限)总体的特征也可以通过一些和概率相联系的总体参数来进行研究. 如果抽样调查为在有限总体(总体量为N)中的不放回随机抽样(样本量为t), 即每个个体有相同的机会被抽到, 而人们感兴趣的是该总体中有某种特征的个体数(未知的m), 那么这种抽样可以用超几何分布(其总体参数为N, t, m)来描述, 这里有限总体的感兴趣的个体数m和超几何分布的总体参数m就是一致的了. 对该有限总体的感兴趣个体的数目m, 可以用对超几何分布的总体参数m的推断来进行研究. 当然有了m, 就有了**总体比例(population proportion)** $p = m/N$(如果对总体比例感兴趣). 如果抽样调查的总体很大, 随机抽样(样本量为n)可以看成为不放回抽样, 如果人们感兴趣的个体在有限总体中的比例是(总体比例)p. 这时抽样可以用总体参数为n和p的二项分布来近似描述. 这时, 有限总体感兴趣的个体的总体比例p和二项分布的总体参数p就是一致的了. 对二项分布参数(试验"成功"概率)p的推断和对总体比例p的推断就一致了.

2.7　习题

1. 举出一些观测数据和试验数据的例子.

2. 举出简单随机样本的例子.

3. 根据你的经验, 举出总体和样本的一些具体例子.

4. 举出调查抽样时可能发生的各种影响调查结果的问题, 并且提出你认为可以减少或避免这些问题的建议.

5. 根据你的直觉, 本章提到的几种抽样方法的优缺点是什么? 原因何在?

6. 举出一些书报上发表的数据例子, 并指出哪些是变量, 哪些是观测值.

第三章 数据的描述

当见过一个人之后，你首先对这个人的外表有个印象，比如高矮胖瘦等等，但更详细的也许一时说不出来. 然而，当你再看到这个人或者这个人的照片时，会马上又认出来. 这说明你的大脑中对这个人除了高矮胖瘦之外还储存了一些其他的信息，只不过一时难以用语言描绘而已，这些信息并不是这个人的全部信息，但能够反映出其某些关键特征. 一大堆数目本身也往往会使人眼花缭乱. 没有人能够记住那些巨大的数据中的所有数值，但总是可以对数据形成一些印象. 有些特征大略了解一下就可以得到. 比如，这些数值的大致范围，是定性还是定量的，有多少变量，以及收集该数据的目的等等. 实际上，借助于一些图形和简单的运算，还可以了解一个数据的更多的特征. 本章介绍如何用简单图表和少数的一些数字来概括数据的某些特征. 当然，由于数据是从总体中产生的，其特征也反映了总体的特征. 对数据的描述也是对其总体的一个近似的描述.

3.1 如何用图来表示数据?

人们对各种图表并不陌生，在中小学时，就可能接触到各种关于成绩、某项活动的进度或者国家发展的各种曲线和图表等等. 在电视、报刊和网络上，也经常有表现股票行情和走势以及其他经济和社会活动的图形. 这些都是统计图形. 下面仅仅把你们可能已经见过的一些统计图形做更详细的解释，可能还会给它们起些专业名字. 这些细节并不用记，只要能够理解图形的意义并会使用计算机软件画出你需要的图形就行了.

3.1.1 定量变量的图表示：直方图、盒形图、茎叶图和散点图

1. 直方图

直方图(histogram)是定量变量最常用的图表示之一. 其作法是，把横轴分成若干通常是等宽度的区间，然后计算数据在各个区间上的频数，并在各区间上画出高度与数据在相应区间的频数成比例的矩形条. 纵坐标当然也可能是比例，即把频数除以样本量，而不一定是频数，但这并不改变图的形状，而仅仅造成纵坐标单位的不同.

例3.1 (数据: Billianaires.txt) 该数据为福布斯(Forbes)公司根据直至2012年3月的资料提供的净资产超过10亿美元的世界富豪排行榜[1]，展示了世界最富有的1223个人(有的包括家族)的名次(Rank，为正整数，越小越富)、名字(Name)、净资产(Net.Worth，单位10亿美元)、年龄(Age)、资产来源(Source，有关行业等信息)、国籍(Country.of.Citizenship). 也就是说有6个变量，其中名次、净资产、年龄为定量变量；名字、资产来源及国籍为定性变量. 我们暂时关心资产和年龄这两个变量. 我们可以用直方图来表示这两个变量的

[1]网址为http://www.forbes.com/billionaires/list/.

数据, 使人们能够看出这些数目的大体分布或"形状". 图3.1就是利用这个数据(Billianaires.txt)由R软件所画的关于这些富人的年龄和财富的两个直方图. 这两个直方图是用下面R语句画出的(第一行读入数据):

```
v=read.table("Billianaires.txt",sep=",",header=T,na.strings="-")
par(mfrow=c(1,2))  #准备画两个并排的图,c(1,2)表示一行两列
hist(v$Age,main="",xlab="Age")
hist(v$Net.Worth,main="",xlab="Net Worth")
```

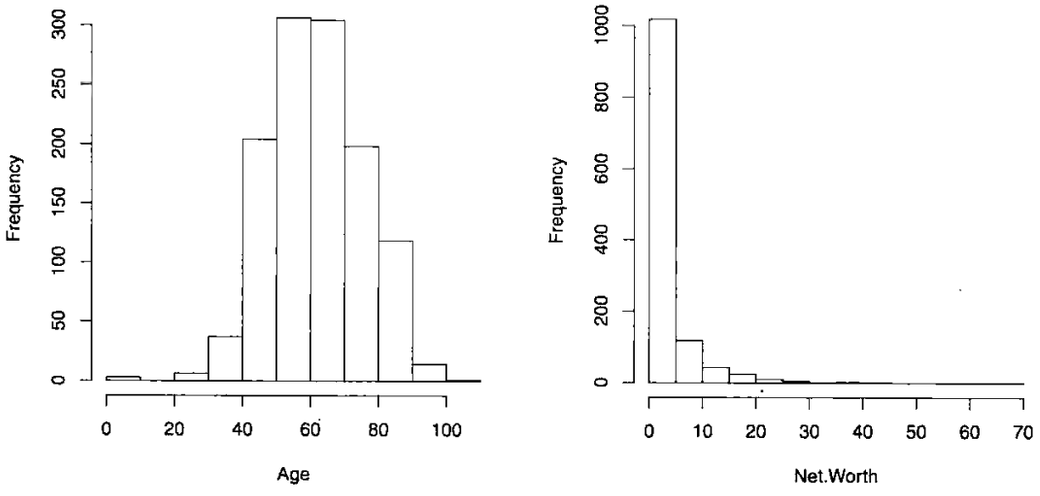

图 3.1　数据Billianaires.txt的富豪的年龄(左)和净资产(右)的直方图.

图3.1左图的横坐标是年龄区间, 这里每一格代表10岁范围(格子宽度因不同的数据性质或要求而定, 这里的格子宽度为10岁), 而纵坐标为各种年龄区间的人数(频数). 在10岁以下的富豪只有3个, 没有10至20岁的富豪. 一些富豪的年龄未知, 因此该直方图没有反映. 右边的关于净资产的直方图(横坐标每一格范围为10亿美元)就不像年龄那样对称了, 大部分人的财富都在50亿以下(一千多个, 左边最高的一个矩形条). 显然从直方图可以看出数据分布的疏密. 显然, 把横轴划分为若干区间有很多选择. 比如, 区间较少时, 则图形只有几个矩形, 而当区间很多时(但相应于数据量还算小时), 则可能会有参差不齐的许多矩形. 确定区间划分的各种方法超出了本书范围. 不过, 各种软件都有一个计算区间的缺省公式. 如果没有把握, 就按照软件的默认方法划分就行了.

2. 盒型图

比直方图简单一些的是**盒形图(boxplot)**, 又称**箱图、箱线图、盒子图**. 图3.2为用Billianaires.txt数据所绘在中、日、美三个国家的富人年龄的盒形图. 该盒形图是用下面R语句画的:

```
w=v[v[,6]=="United States"|v[,6]=="China"|v[,6]=="Japan",]
```

```
w[,6]=as.character(w[,6])
boxplot(Age~Country.of.Citizenship,w)
```

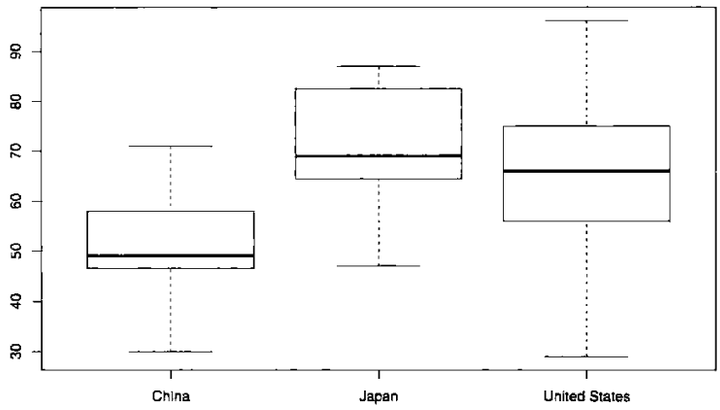

图 3.2 中、日、美三国富豪的年龄的盒形图.

每个盒子的中间的粗横线是数据的**中位数(median)**, 它是下节要引进的量之一. 顾名思义, 中位数是数据中占据中间位置的数, 即数据中约有一半大于中位数(在其之上), 另约一半小于中位数(在其之下). 在把数目按照大小顺序排列之后, 如果数据量为奇数, 那么中间的那个就叫做样本中位数, 如果数据量为偶数, 则中间两个数目的算术平均定义为样本中位数. 封闭盒子的上下两边(横线)为上下四分位数(点)(也是下节要引进的量), 其意义为: 数据中约有四分之一的数目大于上四分位数, 即在盒子之上, 另外有约四分之一的数目小于下四分位数, 即在盒子之下. 因此有一半的数目在中间封闭盒子的范围内. 有一半分布在盒子上下两边. 在盒子上下两边分别各有一条纵向的线段, 表明盒子外面点的分布, 在该线段的两个端点, 各有一条小横线, 标出了最大或最小值. 盒形图可以有不同的画法, 这里是其中一种. 有时, 把离开盒子较远的点单独标出. 此外盒形图可以横过来画, 这都由画图语句的选项控制. 若干个盒形图往往放在一个图中比较. 从该图可以看出中国富人整体上年龄较轻.

3. 茎叶图

在上面介绍的直方图和盒形图中, 已经看不到数值了, 因此很难从图形恢复数据的原貌. 下面引进另一种图: **茎叶图(stem-and-leaf plots)**. 以例3.1数据中的中国富人的年龄为例, 下面的茎叶图是用语句

```
stem(v[v[,6]=="China",4])
```

绘出. 它既展示了年龄的分布形状又有原始数据. 它象一片带有茎的叶子. 茎为较大位数的数字, 叶为较小位数的数字. 可以看出, 该图是用代码打印出来的若干行数目字, 所以不像真正意义上的图形. R软件打印出来的第一行是说明, 指出小数点相应于茎叶界限"|"的位置. 对于这个图, 小数点位于符号"|"往右边一个数字.

```
The decimal point is 1 digit(s) to the right of the |

3 | 0
3 | 899
4 | 0001123344
4 | 55555666667777888888888888889999999
5 | 00111223344
5 | 5556677777788999
6 | 0000012233
6 | 55667789
7 | 01
```

这个茎叶图中茎的单位为10岁, 而叶子单位为1岁. 在第一行指出了一个年龄30岁的. 而第二行茎为30, 因此叶子中的三个数字899代表三个年龄38、39、39岁. 最后一行展示的两个年龄为70和71岁. 显然, 茎叶图既表示了原始数据, 也有直方图显示数据分布的特点. 这有方便的地方. 但是茎叶图有其弱点, 即当数据量很大时(成千上万个度量), 茎叶图就无法显示了. 这也是这里只用了中国富人的数据, 而没有把所有排行榜的富人年龄都画入的原因. 另外, 也可以把几个茎叶图画在一起进行对比. 比如, 两个说明不同总体同样变量的数据可以共用一个茎, "背靠背"地展示叶子, 用来形象地进行比较. 茎叶图并不漂亮, 外行不一定能够马上理解, 因此在媒介中很少出现. 茎叶图显然是前计算机或早期计算机时代的产物.

4. **散点图**

前面的每张图没有显示数量变量之间的关系, 如果需要, 则可以用**散点图(scatter plot)**来描述两个(甚至多个)数量变量的关系. 对于两个变量来说, 在图中, 每一个点代表一个观测值, 而它的横坐标和纵坐标则分别代表其相应于两个变量的取值. 也可以把若干个变量都用纵坐标表示.

例3.2 (数据: global2000.txt, gl100.txt) 该数据为福布斯(Forbes)公司根据2012年4月之前的资料发布的世界上最大的2000个公司的排行榜[1]. 其中数据global2000.txt为全部名单, 而g100.txt为其前100名. 变量包括公司的名次(Rank, 整数, 越小名次越靠前)、公司名称(Company)、所在国家(Country)、销售额(Sales, 单位10亿美元)、利润(Profit, 单位10亿美元)、资产(Assets, 单位10亿美元)、市场价值(Market.Value, 单位10亿美元).

图3.3为福布斯前100名公司的资产(Assets, 横坐标), 销售额(Sales, 纵坐标)和取对数之后的利润(log Profits, 体现在符号的大小)的散点图. 该图还标出了销售额最大的6个公司(Royal Dutch Shell, Wal-Mart Stores, Exxon Mobil, Sinopec-China Petroleum, BP, PetroChina)以及资产最多的公司(Deutsche Bank).

绘制图3.3的R代码为(其中`identify()`为互动式的手工选择函数):

[1]该数据网址为http://www.forbes.com/global2000/list/.

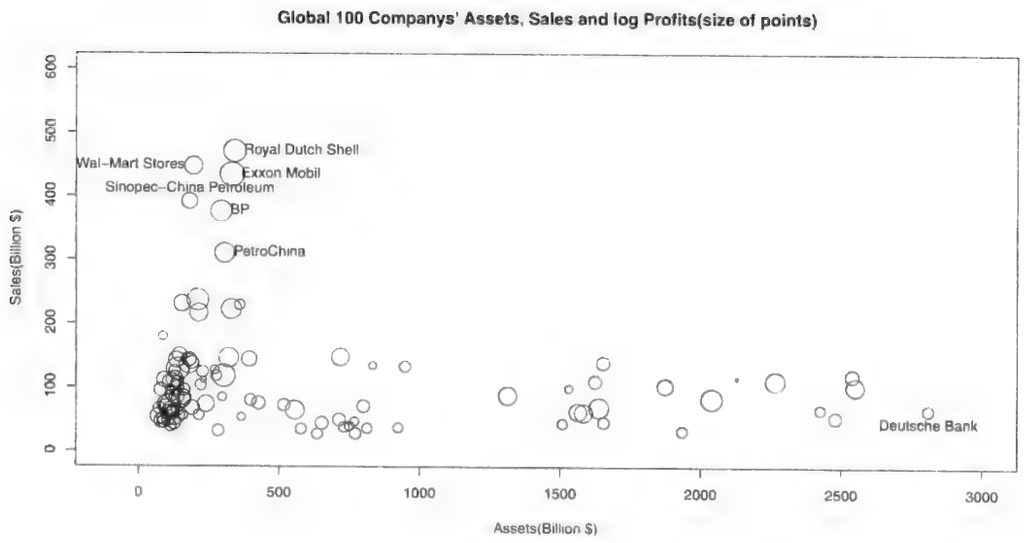

图 3.3 展示福布斯前100名公司的资产、销售额和利润(对数)的散点图.

```
v=read.table("g100.txt",sep=",",header=T)
plot(v$Assets, v$Sales,pch=1,col=1,xlab="Assets(Billion $)",
ylab="Sales(Billion $)",ylim=c(0,600),xlim=c(-100,3000),cex=log(v$Profits))
title("Global 100 Companys' Assets, Sales and log Profits(size of points)")
identify(v$Assets, v$Sales,labels=v$Company)
```

3.1.2 定性变量的图表示：饼图和条形图

定性变量(或属性变量, 分类变量)不能点出直方图、散点图或茎叶图, 但可以用图来描绘出它们各类的数目或者其他数量特征的比例. 还是用例3.2来说明.

1. 饼图

饼图(pie chart)为一个由许多扇形组成的圆, 各个扇形的大小比例等于变量各个水平(类)的频数或者是相关数量变量的比例. 饼图比条形图简单, 描述比例较直观. 但是当变量太多时饼图就不那么好看了. 图3.4表示包含大公司最多的国家中的前10名的公司数目的饼图. 该图的代码(包括读入数据)如下:

```
w=read.table("global2000.txt",sep=",",header=T)
ws=sort(table(w$Country),de=T);pie(ws[1:10])
title("Number of Companies Among top 10")
```

2. 条形图

从图3.4的饼图中仅仅看出了各个国家拥有的大公司的比例, 而看不出具体数目, 为这个目的, **条形图(bar plot)**就更适当了. 图3.5为用图3.4同样数据所绘的条形图. 该图的代码如下:

```
barplot(ws[1:10],cex.names =.8,main="Number of Companies Among top 10")
```

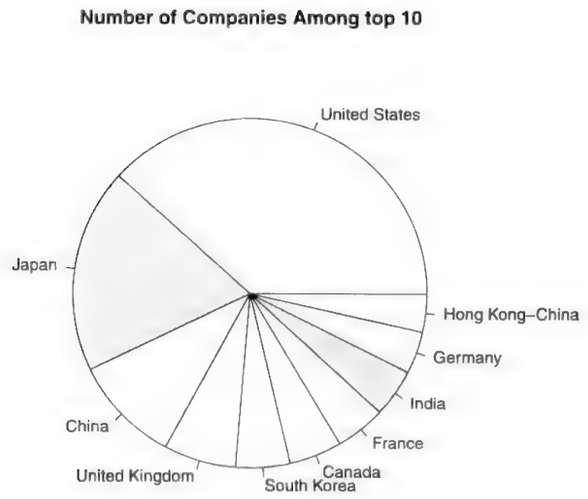

图 3.4　包含公司最多的国家中的前10名的公司数目的饼图.

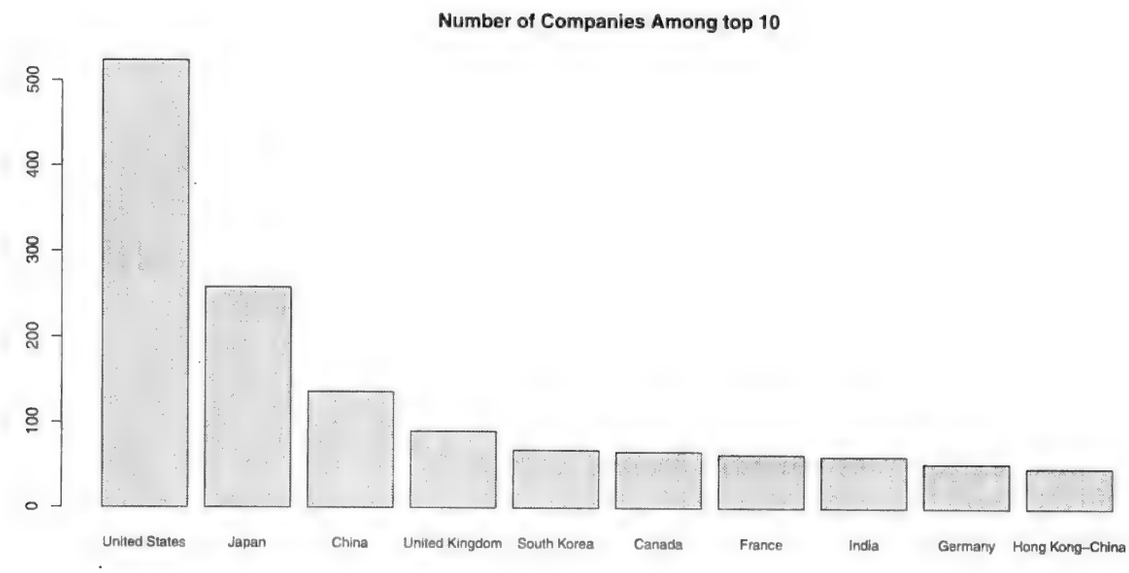

图 3.5　包含公司最多的国家中的前10名的公司数目的条形图.

3.1.3 其他图描述法

除了上面说的各种用来描述数据的图之外, 还可采用其他各种图形. 下面介绍其中几种.

1. Chernoff 面孔图和星图

有一种很独特的**Chernoff 面孔图(Chernoff Faces)**, 它把矩阵形式的数据用面孔形式表现出来. 不同的面孔体现数据各个变量的不同特征. 当然, 你必须熟悉这些面孔的各种器官和表情代表数据的什么特征才行. 各个变量相应的器官度量包括面孔长度、面孔宽度、面孔形状、嘴的上下高度、嘴的宽度, 笑容的曲线、眼睛的睁开程度、眼睛的宽度、头发的厚度、头发的宽度、鼻子的长度、鼻子的宽度、耳朵的宽度、耳朵的长度等. 各种变量的组合就形成面孔的不同表情. 另一种图为**星图(star plot)**, 也称**蜘蛛**或**雷达图(spider/radar plot)**, 它把各个变量按照大小向各个方向做射线段, 形成星辰形状. 这个图比面孔图容易理解, 但比较死板.

图3.6和图3.7为用销售额、利润、资产及市场价值4个变量来描述世界前10位的大公司的Chernoff面孔图(图3.6)和星图(图3.7). 这两个图利用了程序包`TeachingDemos`[1], 是用下面语句得到的:

```
v=read.table("g100.txt",sep=",",header=T)
library(TeachingDemos);
q=v[1:10,4:7];row.names(q)=v[1:10,2]
faces(q,nrow=2,ncol=5);stars(q,nrow=2,ncol=5)
```

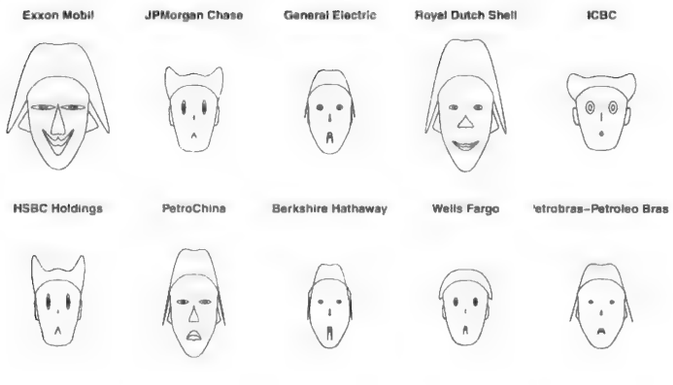

图 3.6 世界前10位的大公司的Chernoff面孔图.

[1]Greg Snow (2012). TeachingDemos: Demonstrations for teaching and learning. R package version 2.8. http://CRAN.R-project.org/package=TeachingDemos.

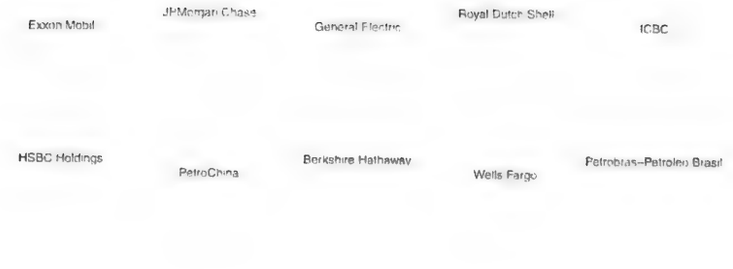

图 3.7　世界前10位的大公司的星图.

2. Lorenz曲线

Lorenz曲线和Gini系数不属于本书范围, 但有些经济背景的读者应该了解这一对概念. Lorenz曲线的横坐标为从最低收入到最高收入的人口的累积比例(从0%到100%), 其纵横坐标为人们挣得的收入从最低到最高的累积份额(从0%到100%). 如果人们收入全一样, 那么Lorenz曲线应该是45度直线; 如果是向下凸的曲线, 那么该曲线和45度对角线线之间的面积越大, 则说明越不平等. Gini系数为该面积和整个三角形面积之比. 例3.1数据的1223个富人共有45660.3亿资产. 但就这些资产而言, 他们之间的差距如何呢? 我们可以画出Lorenz曲线, 计算出该数据的Gini系数为0.4877021. 这说明富人世界中也是"贫富不均"了. 当然, 真正的Lorenz曲线应该用一个国家或一个地区的所有人的收入来计算. 这里仅仅是借用富人的资产数据来介绍Lorenz曲线. 该曲线在图3.8之中. 绘图利用了程序包`ineq`[1], 代码为:

```
v=read.table("Billianaires.txt",sep=",",header=T,na.strings="-")
library(ineq); plot(Lc(v[,3]),col='red');Gini(v[,3])
```

思考一下:

1. 用什么图形来描述各个月度的GDP和物价CPI之间的关系?
2. 用什么图形来描述不同教育程度的人口的比例(假定分小学、中学、大学、研究生以上)?
3. 用什么图形来描述一个年级身高的度量?

[1]Achim Zeileis (2012). ineq: Measuring Inequality, Concentration, and Poverty. R package version 0.2-10. http://CRAN.R-project.org/package=ineq.

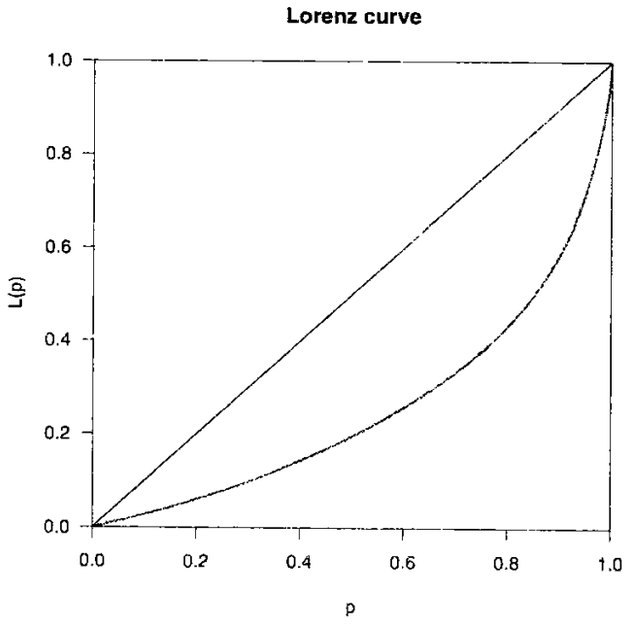

图 3.8 例3.1数据按照资产的Lorenz曲线.

3.2 如何用少量数字来概括数据?

用少数几个数字概括大量数字是日常生活中常见的. 比如说, 北京人的平均收入是多少、两地区的收入差距是多少、高收入的人占人口的百分比等. 这些"平均"、"差距"或百分比都是用来概括或汇总的数字. 由于定性变量主要是计数, 比较简单, 其常用的概括就是比例或百分比, 所以下面主要介绍关于定量变量的数字描述.

除了图表之外, 可以用少量所谓**汇总统计量或概括统计量(summary statistic)**来描述定量变量的数据. 这些数字是从样本得来的, 因而也是样本的函数, 任何样本的函数, 只要不包含总体的未知参数, 都称为**统计量(statistic)**. 样本本身是随机的, 从同一个总体抽出来的不同样本肯定不一样. 因此, 对于不同数据(即样本的实现), 统计量的取值也不一样, 也就是说样本的随机性决定了统计量的随机性.

在许多情况, 从样本产生的一些统计量的实现值反映了无法观测到的某些总体参数的大小, 这时统计量就可以用来作为这些参数的估计. 以后还要提到, 作为样本某种代表的一些统计量还可以用来检验样本和假设的总体是否一致. 一些统计量前面有时加上"样本"二字, 以区别于总体的同名参数. 比如后面的从样本产生的均值和标准差严格说来应该叫做"样本均值"和"样本标准差", 以区别于总体的均值和标准差. 但在不会混淆时可以只说"均值"和"标准差". 一些总体参数将在下一章介绍.

3.2.1 数据的“位置”

人们常说哪个地方穷, 哪个地方富. 也常说, 哪个国家人高, 哪个国家人矮. 说这些话的人绝对不是说一个地方的所有人都比另一地方的所有人富, 也不是说, 一个国家的人都比另一个国家的所有人都高. 他们仅仅省略了“平均起来”, “大部分”等词语. 这些说法实际上是关于数据中某变量观测值的“中心位置” 或者数据分布的中心(center或center tendency)的某种表述. 这种与“位置”有关的统计量就称为**位置统计量(location statistic)**. 位置统计量当然不一定都是描述“中心”了, 比如后面要讲的k百分位数.

最常用的位置统计量就是小学时所学到的算术平均值, 它在统计中叫做**均值(mean)**, 严格地说叫做**样本均值(sample mean)**, 以区别于下一章要介绍的总体均值. 样本均值是把一个变量的所有观测值求和再除以观测值的数目. 如果记样本中的观测值为$x_1, ..., x_n$, 则样本均值定义为

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + \cdots + x_n}{n}.$$

比如, 可以用上面公式得到例3.1中富豪的平均年龄. 由于只有1191个富豪的年龄已知, 所以$n = 1191$. 利用R代码`mean(v$Age,na.rm=T)`可得均值为62.4岁. 公式中的选项`na.rm=T`表示去掉缺失值(即不知道的年龄)再求平均.

虽然均值包含了样本的很多信息, 但它容易被少数极端值所影响. 比如, 一个数据输入员的疏忽很可能造成某些数目出错, 比如多敲入若干0, 这时均值就可能变成很大. 但这种数据错误不会对该数据按升幂或降幂排列的中间一些数目影响太大. 数据中间的一个(或两个数的平均)就是**(样本)中位数(median)**. 它是数据按照大小排列之后位于中间的那个数(如果样本量为奇数), 或者中间两个数目的平均(如果样本量为偶数). 利用R代码`median(v$Age,na.rm=T)`可得例3.1富豪年龄的中位数是62岁. 由于中位数不易被极端值影响, 所以称中位数比均值**稳健(robust)**. 比如一千个月收入为2000元的人和月收入为一千万元的一个富翁住在同一个区域, 则该区域人们的“平均”月收入用均值计算为11988.01元, 而用中位数计算为2000元, 相差将近6倍.

描述数据“中心位置”的均值、中位数各有优缺点. 但也有一定的规律. 对于具有对称单峰分布(“对称”相应于对称直方图所反映的形状, 所谓“单峰”是分布中只有一个体现局部极大值那样的“峰”), 这两个度量应该大体上差不多. 而如果单峰的分布形状在右边拖尾(即直方图在右边有长尾巴), 那么一般说来, 中位数小于均值. 反过来, 如果直方图在左边拖尾, 则一般均值小于中位数. 也就是说, 和中位数相比, 均值一般总是在长尾巴那边.

中位数在数据大小顺序中居中. 而前面提到的上下**四分位数**(或分别称为**第一四分位数**和**第三四分位数, first quantile, third quantile**) 则分别位于(按大小排列的)数据的上下四分之一的地方. 一般地还称上四分位数为**75百分位数(75 percentile,** 有约75%的观测值小于它), 下四分位数为**25百分位数**(有约25%的观测值小于它). 有了25百分位数和75百分位数的概念, 人们就不难理

解什么是任意的k百**分位数(k-percentile)**了(有约$k\%$的观测值小于它). 如果令$\alpha = k\%$, 则k百分位数也称为**α分位数(α-quantile)**. 显然中位数是50百分位数或0.5分位数. 根据例3.1数据, 富人的两个四分位点分别为52和72岁, 可用语句`quantile(v$Age,.25, na.rm=T)`和`quantile(v$Age,.75, na.rm=T)`计算.

除了中位数和均值之外, 还有样本中出现最多的某一数目, 称为**众数(mode)**. 例3.1富豪榜数据中富豪年龄的众数为60岁, 一共有41位这个年龄的人(用代码`z=table(v$Age);z[which(z==max(z))]`得到). 注意, 如果年龄精确到分钟甚至秒, 则不大可能会有众数. 众数反映的信息也不多, 又不一定唯一, 在连续变量的情况, 如果不做过分四舍五入, 可能没有重复的数据, 这时也不可能有众数. 众数用得不如均值和中位数普遍. 在定性变量中, 由于记录的是频数, 因此众数用得多些. 比如在图3.5关于10个国家拥有大公司的数目的条形图中, 众数就是由美国所代表, 它一共拥有524家大公司.

3.2.2 数据的"尺度"

论语有一句话: "不患寡而患不均". 这是指不怕财富少, 而怕分配不公平而造成贫富差距太大. 贫富、多寡是由位置统计量来描述的, 而是否"均"是由**尺度统计量(scale statistic)**来描述的. 尺度统计量是描述数据散布, 即描述集中与分散程度或变化(spread或variability)的度量, 因此, 有人不无道理地建议用"散度统计量"这个名词. 统计中有许多尺度统计量. 一般来说, 数据越分散, 尺度统计量的值越大. 为了说明, 回顾图3.2, 那是中国、日本和美国富豪的年龄的盒形图, 可以看出, 从中位数来说, 日本年龄较大, 美国次之, 中国最小, 分别为69, 66和49. 而均值分别为69.1, 65.5和52. 但是这三个数据散布范围和模式很不一样.

最简单的尺度统计量就是**极差(range)**, 顾名思义, 极差就是极大值和极小值之间的差. 例3.1数据的中日美三国富豪年龄的极差分别为41, 52, 和69岁. 图3.2中每个盒形图盒子的长度为上下两个四分位数之差, 称为**四分位数极差**或**四分位间距(interquantile range)**, 它描述了中间半数观测值的散布情况. 极差和四分位极差实际上各自只依赖于两个值, 信息量太少. 例3.1的中日美三国富豪年龄的四分位极差分别为11.5, 16.5和19.

另一个常用的尺度统计量为**(样本)标准差(standard deviation)**. 它度量样本中各个数值到均值的距离的一种平均. 标准差实际上是**方差(variance)**的平方根. 样本方差是由各观测值到均值距离的平方和除以减去1的样本量. 也就是说, 如果记样本中的观测值为$x_1, ..., x_n$, 则样本方差为

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}.$$

而样本标准差为样本方差的平方根:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}}.$$

标准差由于和原数据量纲一样, 因此在数据分析中比方差用得更普遍.

显然, 如果标准差越大, 数据中的观测值就越分散, 而小的标准差意味着数据很集中. 下一章会介绍总体标准差和总体方差的概念. 关于中日美三国富豪的标准差分别是8.4, 15.1和13.0.

在直方图中只有一个最高点的数据被称为单峰的, 如果还左右对称, 则是单峰对称数据. 服从(下一章要介绍的)正态分布的数据就是单峰对称的. 对于正态分布的数据, 均值左右一个标准差的范围应该会包含大约68%的观测值, 而均值左右两个标准差的范围应该会包含大约95%的观测值, 均值左右三个标准差的范围应该会包含大约99.7%的观测值(也就是绝大部分观测值). 一些人把这种粗略的准则推广到一般单峰对称数据上, 这时这种经验法则必然会和实际情况有出入, 而相差多少则依赖于具体数据的性质.

即使出于同一个总体, 样本量相同的不同样本也会有不同的均值, 这种来自许多不同样本的均值的标准差称为**标准误差(standard error)**, 也叫做**均值的标准误差(standard error of mean)**. 样本均值的各种性质包括大样本分布性质可参看第四章的抽样分布和中心极限定理部分. 由于不同样本所产生的均值比一个样本中的观测值要稳定得多, 它的标准差比针对整个数据的标准差要小得多. 标准误差定义为标准差除以样本量的平方根, 即$s/\sqrt{n}$.

3.2.3 数据的标准得分

例3.3 (数据: grade.txt) 该数据给出两个班(一班和二班)的同一门课的成绩. 假定两个班水平类似, 但是由于两个任课老师的评分标准不同, 使得两个班成绩的均值和标准差都不一样. 一班分数的均值和标准差分别为78.53和9.43, 而二班的均值和标准差分别为70.19和7.00. 那么得到90分的一班的张颖是不是比得到82分的二班的刘疏成绩更好呢? 怎么比较才能合理呢?

虽然这种均值和标准差不同的数据不能够直接比较, 但是可以把它们进行标准化, 然后再比较标准化后的数据. 一个标准化的方法是把某样本原始观测值(亦称得分, score)和该样本均值之差除以该样本的标准差, 得到的度量称为**标准得分(standard score**, 又称为**z-score**). 即, 某观测值x_i的标准得分z_i定义为

$$z_i = \frac{x_i - \bar{x}}{s}.$$

把各个样本的观测值都转换成相应的标准得分, 就可以进行比较了. 在这个例子中, 张颖的标准得分为$(90 - 78.53)/9.43 = 1.22$, 而刘疏的标准得分为$(82 - 70.19)/7 = 1.69$. 显然如果两个班级水平差不多, 刘疏的成绩应该优于张颖的成绩, 这是在标准化之前的数据中不易看到的.

图3.9展示了这两个班级的原始成绩的盒形图(左边)和标准化之后成绩的标准得分的盒形图(右边). 可以看出, 原始数据是在各自的中心值附近, 而散布也不一样. 但它们的标准得分则在0周围散布, 而且散布也差不多. 实际上, 任何样本经过这样的标准化后, 就都变换成均值为0、方差为1的样本. 标准化后不同样本观

测值的比较只有相对意义, 没有绝对意义.

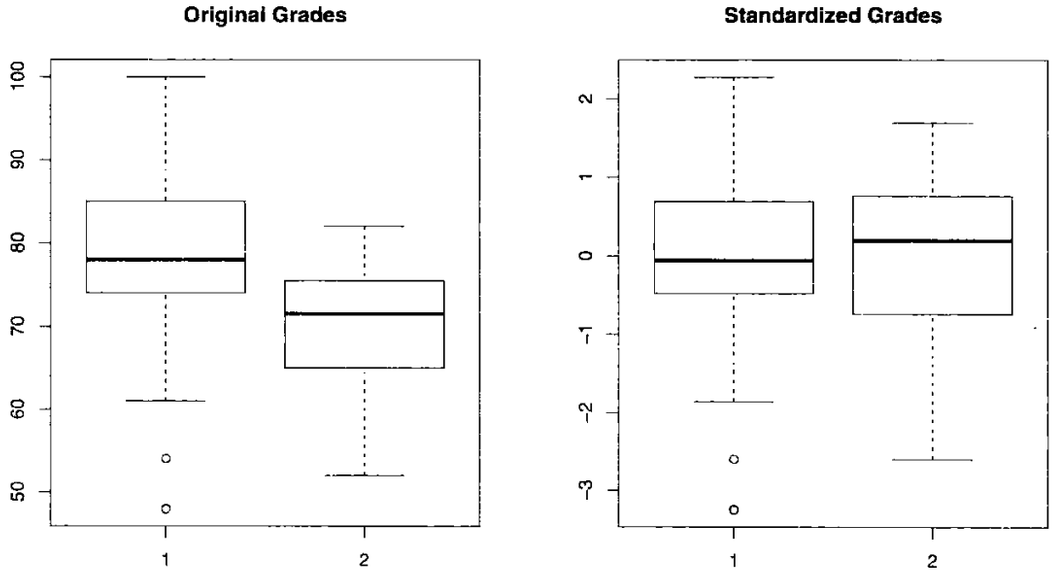

图 3.9 例3.3数据两个班级分数的原始数据(左)和标准得分(右)的盒形图.

绘出图3.9的R代码为

```
w=read.table("grade.txt",header=T)
par(mfrow=c(1,2))
boxplot(grade~class, w,main="Original Grades")
boxplot(standardized~class, w,main="Standardized Grades")
```

标准化之后的数据虽然总的尺度和位置都变了, 但是数据内部点的相对位置没有变化. 比如, 距离均值两倍标准差的一个点在标准化后距离均值还是两倍标准差. 这从图3.9也可以看出: 每个数据标准化前和标准化后的盒形图(在纵向)相似. 这是因为标准化仅仅是把盒形图进行纵向放大(或缩小)和位移. 班级1的两个离群点还是离群点. 虽然如此, 但两个不同的数据在标准化后就有了进行比较的基础. 标准得分的思想不仅仅用于比较, 而且在后面的推断中也有其用处. 另外, 计算标准得分也仅仅是许多标准化方法中最常见的一种.

无论问题是什么, 一些人喜欢把各种数据都标准化之后再进行分析, 这是不适当的. 把例3.3数据标准化的前提是两个班级是相似的. 如果对于完全不同背景的数据进行标准化, 就会失去很多有用的信息. 比如, 把最富有的国家和把最贫穷的国家的收入进行标准化, 结果的标准化后的数据有同样的均值0及同样的标准差1, 结果是, 基本上看不出哪个数据是来自哪个国家. 很难说这样的标准化除了误导之外还有什么意义.

当然, 在应用一些统计方法时, 有时的确需要对数据做标准化或其他变换, 但这些并不是随意的, 都有某些确定的理论基础和实际目的.

思考一下:

1. 前面描述位置统计量上下四分位点时, 说"约75%的观测值小于它"和"约25%的观测值小于它", 为什么要说"约"呢? 不能精确点吗? 这是因为数据点的个数是离散的缘故. 为了理解, 请大家找出由三个点(比如1、2、3)组成的数据的上下四分位点, 并且看有百分之几的样本点小于下四分位点.

2. 为什么说中位数比均值稳健? 有一种截尾均值, 定义为把数据的最大的及最小的一定比例的数目去掉之后再求均值, 这可以用R实现. 比如mean(x,trim=0.1)就是把数据x的最高和最低两个尾巴各去掉10%的数据再求算术平均, 这里trim取值从0到0.5. 当trim=0.5时, 就得到中位数, 请解释.

3. 为什么说, 原则上连续变量的数据不应该有众数存在, 但实际上可能会出现? 是不是四舍五入把连续变量的实现值的记录离散化了? 实现值在记录之前是否应该有众数?

4. 一个样本的位置统计量和尺度统计量在求样本点的标准得分前后有什么变化. 一个班级内部的学生名次是否因为求了标准得分之后会改变?

3.3 小结

3.3.1 本章的概括和公式

本章涉及如何用图和少量数字来描述数据. 对于定性变量来说, 有饼图和条形图, 而对于定量变量, 有直方图、茎叶图、盒形图和散点图等. 当然这些图仅仅包含最常用的那些图. 除了图表示之外, 定量变量的数据还可以用少数几个数来描述该数据的位置(位置统计量), 这包括描述数据"中心位置"的众数、均值和中位数, 以及描述极端值及其他位置的百分位数. 定性变量的汇总统计量包括百分比及"众数"(百分比最大的那一类). 另外还介绍了描述定量变量尺度, 即数据散布(或集中)程度的统计量, 它们有极差, 标准差, 方差, 四分位极差等等. 对于样本均值的标准差, 引进了标准误差. 为了比较不同均值和不同方差的数据点, 本章还介绍了标准化的方法, 即用标准得分代替原先的数据来比较.

定义和公式

假定某样本的**样本量(sample size**, 即样本中观测值的个数)为n, 样本中的观测值为$x_1, ..., x_n$, 则样本均值定义为

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + \cdots + x_n}{n}.$$

假定观测值按照自小到大的升幂排列为$x_{(1)}, ..., x_{(n)}$, 则当n为奇数时**样本中**

位数定义为$x_{((n+1)/2)}$，而当n为偶数时样本中位数定义为$[x_{(n/2)}+x_{(n/2+1)}]/2$.

样本方差定义为

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{(x_1-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n-1}.$$

而样本标准差为样本方差的平方根:

$$s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}=\sqrt{\frac{(x_1-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n-1}}.$$

样本标准误差为

$$s.e.(x)=\frac{s}{\sqrt{n}}=\frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}}{\sqrt{n}}.$$

某观测值x_i的标准得分为

$$z_i=\frac{x_i-\bar{x}}{s}.$$

3.3.2 R语句的说明

1. 做图形

下面假定x是数据向量.

- 直方图: `hist(x)`
- 盒形图: `boxplot(x)`
- 茎叶图: `stem(x)`
- 散点图: `plot(x)`
- 饼图: `pie(x)`
- 条形图: `barplot(x)`

2. 计算汇总统计量、位置、尺度、标准得分等

- 汇总统计量: `summary(x)`
- 均值: `mean(x)`
- 中位数: `median(x)`
- 分位点: `quantile(x)`或`quantile(x,0.75)`等

- 极差: `diff(range(x))`
- 四分位极差: `diff(quantile(x,c(.25,.75)))`
- 标准差: `sd(x)`
- 方差: `var(x)`
- 标准得分: `scale(x)`

3.4 习题

1. 根据你的经验, 给出定性和定量变量的例子, 并试图画出各种描述性图形并计算汇总统计量.
2. 举例说明众数、中位数和均值的优缺点.
3. 尺度统计量说明了数据的什么特性? 举例说明.
4. 标准得分实际上是对原始数据的一种标准化. 试举出标准得分的用处. 何时不能做标准化?

第四章　机会的度量: 概率和分布

前面已经提到, 概率是0和1之间(包含0和1)的一个数目, 表示某个事件发生的可能性或经常程度. 有些事情发生的概率很大, 而有些则很小. 比如, 你乘车出门可能遇到车祸的概率很小(也许几乎是0), 但在北京一天中发生一起以上的车祸的概率几乎是1. 你今年买房子的概率可能很小, 但在北京每天有人买房子的概率却很大. 发生概率很小的事件称为**小概率事件(small probability event)**, 虽然小概率事件不那么可能发生, 但它往往比很可能发生的事件更值得研究. 这一章介绍如何得到概率、如何进行概率计算以及什么是概率分布, 还要介绍一些常用的分布. 许多读者可能已经熟悉这一章的许多内容, 对于他们, 这一章可以跳过. 由于本章的图形是为了帮助读者理解概念而绘制的, 画图方法本身与这些概念无关, 为了不影响内容的连贯性, 我们把生成这些图形的R代码放到后面4.6.3节, 供感兴趣的读者参考.

4.1　得到概率的几种途径

利用等可能事件

如果一个骰子是公平的[1], 那么掷一次骰子会以相等可能得到1至6点中的每一个点. 这是因为共有$n = 6$种可能, 而每一种的概率都是一样的, 即$1/n = 1/6$. 抛一个公平的硬币(并假定不可能得到侧面), 则以等可能出现正面或反面. 这是因为只有$n = 2$种可能, 每种概率都是$1/n = 1/2$. 再如从52张牌中随机抽取一张, 那么它是黑桃的概率是抽取黑桃的可能种类($k = 13$)和全部可能种类($n = 52$)的数目之比, 即$k/n = 13/52 = 1/4$, 类似地, 抽到的牌是J、Q、K、A四种之一(共有16种可能)的概率是$16/52 = 4/13$. 其实即使没有学过概率, 读者也多半能够算出这些概率, 这种计算的原理一般都不太困难. 计算这些概率的基础就是事先知道(或者假设)某些事件是等可能的. 这种事件称为**等可能事件(equally likely event)**.

根据长期相对频数

在多数情况下, 事件并不一定是等可能的, 或者人们对于其出现的可能性一无所知. 这时就要靠观察它在大量重复试验中出现的频率来估计它出现的概率. 它约等于事件出现的频数k除以重复试验的次数n, 这个比值k/n称为**相对频数(relative frequency)**或**频率**. 例如, 在估计新生儿中男婴的比例时, 如果在$n = 1000$个新生儿中有$k = 517$个男婴, 那么就可以近似地说, 生男婴的概率为$k/n = 517/1000 = 0.517$. 就这个例子来说, 人们可能会认为, 当n趋于无穷时,

[1]本书中凡是提到"公平的骰子", 或"骰子是公平的", 意味着该骰子是用均匀材料制成的正六面体, 在其六面上分别标有一至六点, 在它被掷出后, 每一面朝上的机会是均等的. 类似地, "公平的硬币", 也意味着该硬币被掷后, 两面以同样机会朝上的硬币, 当然, 公平硬币的两面是不同的, 常称一面是"正面"(比如国徽或花卉), 而另一面为"反面"(比如面值). 而一个灌了铅或水银的骰子或两面相同的假硬币都不是公平的.

这个相对频率趋于真正生男婴的概率. 但是要注意的是, 不可能观测无穷多次出生, 而且真正的生男婴的概率也可能随环境而变. 在商业实践中一个类似的例子是, 假定想知道某个橱窗设计吸引注意力的概率, 可以观察有多少过往的人在它面前逗留观看. 如果观察了$n = 500$人(500次试验), 有$k = 12$个人在该橱窗前逗留, 那么可以大致地说, 该橱窗吸引行人的概率近似地为相对频数$k/n = 12/500$. 试验次数n越大则该值越接近于想得到的概率. 再如, 对某个商品投诉的概率近似地为投诉数目(k)除以售出的该种商品数(n), 即k/n. 这里每卖出一个商品为一次试验. 卖出的商品越多, 则这个比例越接近投诉的概率.

很多事件无法进行长期重复试验, 或者根本不可能重复. 因此这种通过相对频数获得概率的方法并不是万能的. 虽然如此, 用相对频数来确定概率的方法是很常用的.

主观概率

一些概率既不能由等可能性来计算, 也不可能从试验得出. 比如, 你明年去九寨沟旅游的概率、一个公司的董事会是否明天要讨论某个问题的概率等都无法重复. 但根据经验、常识或其他相关因素来判断, 你可能会说出一个概率. 比如你明年去九寨沟的概率是百分之八十等. 这种概率称为**主观概率(subjective probability)**. 可以说, 主观概率是一次事件的概率. 也可以说, 主观概率就是基于对各种信息的掌握, 某人对某事件发生或者对某断言的真实性的自信程度.

思考一下:

1. 在实际生活中, 等可能性事件不易见到, 很难找到各种条件完全相同的实验. 例如, 掷骰子、抛硬币或者洗牌的手法不同, 很可能会造成不公平的结果. 你对此如何看?

2. 讨论使用若干相对频率确定概率的优缺点. 你可以在气象预报、科学实验、战争发生的可能性预测等各种方面来讨论.

4.2 概率的运算

在掷骰子中, 得到6点的概率是1/6, 而得到5点的概率也是1/6. 那么掷一次骰子得到5或者6的概率是多少呢? 在掷两次骰子中两次都得到5或者6的概率又是多少呢? 在掷10次骰子中有一半或以上的次数得到5或6的概率又是多少呢? 读者略微思考一下就可能很快会得到答案. 如果情况再复杂一些, 答案也许就不是那么简单了. 这就需要了解怎样从简单的情况计算稍微复杂情况时的概率. 这里需要读者回忆一下上中学时学过的集合概念, 比如两个集合的交和并, 互余(互补)等概念. 在概率论中所说的**事件(event)**相当于集合论中的**集合(set)**. 而概率则是事件的某种函数. 为什么会这么说呢, 让我们看掷两个骰子的试验. 如果所关心的是两个骰子的点数和, 则下表列出了了所有36种可能试验结果的搭配和相应的点数和. 每次试验结果为其中之一. 从表中可以看出, 如果我们考虑点数和等于2的事

件, 则仅有一种可能的试验结果(两个骰子均为一点). 而如果我们考虑点数和等于7的事件, 则有六种可能的试验结果. 两个骰子点数之和总共有2至12等11种可能, 即有11种可能的事件, 而这11种事件相应于上面所说的36种可能的试验结果的一些集合.

两个骰子之和的试验

事件:骰子点数之和	集合: 相应的试验结果(36种搭配)	集合中元素的个数	事件的概率
2	(1,1)	1	1/36
3	(1,2) (2,1)	2	2/36
4	(1,3) (2,2) (3,1)	3	3/36
5	(1,4) (2,3) (3,2) (4,1)	4	4/36
6	(1,5) (2,4) (3,3) (4,2) (5,1)	5	5/36
7	(1,6) (2,5) (3,4) (4,3) (5,2) (6,1)	6	6/36
8	(2,6) (3,5) (4,4) (5,3) (6,2)	5	5/36
9	(3,6) (4,5) (5,4) (6,3)	4	4/36
10	(4,6) (5,5) (6,4)	3	3/36
11	(5,6) (6,5)	2	2/26
12	(6,6)	1	1/36

注: 试验结果括号中的两个数字分别表示第一和第二个骰子的点数.

上表中的每一行第一列都是骰子点数之和(即事件)的一种(从2到12); 而每行的第二列为产生这个和(事件)的各种可能的试验结果, 这些试验结果形成一些集合(每行一个集合); 每行的第三列为该行集合中元素的个数; 每行最后一列为得到这种骰子和的概率, 由于所有试验结果是等可能的, 所以得到某种和(事件)的概率等于该集合中试验结果的个数除以所有可能试验结果的个数. 这样, 我们就把事件、集合与概率联系到一起了.

下面介绍一些概率的运算.

互补事件的概率

如果今天淋雨的概率是80%, 那么, 今天不淋雨的概率就是20%. 如果这个月中奖的概率是0.0001, 那么这个月不中奖的概率就是$1-0.0001=0.9999$. 这种如果一个不出现, 则另一个肯定出现的两个事件称为**互补事件(complementary events, 或者互余事件或对立事件)**. 按照集合的记号, 如果一个事件记为A, 那么另一个记为A^C(称为A的余集或补集). 显然互补事件的概率之和为1, 即$P(A)+P(A^C)=1$, 或者$P(A^C)=1-P(A)$. 这里记号$P(A)$的英文含义为probability of A. 在西方赌博时常常爱用**优势或赔率(odds)**来形容输赢的可能. 在生物统计中也常用优势的概念. 它是互补事件概率之比, 即$P(A)/P(A^C)=P(A)/[1-P(A)]$来表示的. 如果你赢的概率为0.6, 那么你的优势为$0.6/(1-0.6)=0.6/0.4=6/4$, 说成是你有6对4的优势会赢, 或4对6的优势会输.

概率的加法

如果两个事件不可能同时发生, 那么至少其中之一发生的概率为这两个事件的概率和. 比如"掷一次骰子得到5或者6点"的概率是"得到5点"的概率与"得到6点"的概率之和, 即$1/6+1/6=1/3$. 但是如果两个事件可能同时发生时这样做就不对了. 假定掷骰子时, 一个事件A为"得到偶数点"(有3种可能: 2、4、6点), 另一个事件B为"得到大于或等于3点"(有4种可能: 3、4、5、6点), 这样, 事件A的概率显然等于$3/6=1/2$, 即$P(A)=1/2$. 而事件B的概率为$P(B)=4/6=2/3$. 但是, "得到大于或等于3点或者偶数点"的事件的概率就不是$P(A)+P(B)=1/2+2/3=7/6$了, 这显然多出来了. 概率怎么能够大于1呢? 按照中学时关于集合的记号, 该事件称为A和B的并, 记为$A\cup B$. 刚才多出来的部分就是A和B的共同部分$A\cap B$(称为A和B的交)的概率(这个概率算了两遍), 它为"得到既是偶数, 又大于等于3"的部分, 即4和6两点. 出现事件4或者6的概率为$1/6+1/6=1/3$. 于是应该把算重了的概率减去. 这样"得到大于或等于3点或者偶数点"的事件$A\cup B$的概率就是$P(A\cup B)=P(A)+P(B)-P(A\cap B)=1/2+2/3-1/3=5/6$. 当然, 这个问题也可以换个角度来看, 如果记事件$C=A\cup B$, 那么C^C就是既不是偶数又小于3点的事件, 也就是说只有1点了. 它的概率为$1/6$, 即$P(C^C)=1/6$. 这样根据互补事件的概率, $P(A\cup B)=P(C)=1-P(C^C)=1-1/6=5/6$. 这种$P(A\cup B)=P(A)+P(B)-P(A\cap B)$的公式也适用于两个不可能同时发生的事件, 但因为那时$P(A\cap B)=0$, 所以只剩下$P(A\cup B)=P(A)+P(B)$了. 这种交等于空集($A\cap B=\phi$, 这里ϕ表示空集或空事件)的事件为两个不可能同时发生的事件, 称为**互不相容事件(mutually exclusive events)**.

概率的乘法

如果你有一个固定电话和一个手机, 假定固定电话出毛病的概率为0.01, 而手机出问题的概率为0.05, 那么, 两个电话同时出毛病的概率是多少呢? 聪明的读者马上会猜出, 是$0.01\times 0.05=0.0005$. 但是这种乘法法则, 即$P(A\cap B)=P(A)P(B)$, 仅仅在两个事件**独立(independent)**时才成立.

如果事件不独立则需要引进**条件概率(conditional probability)**. 比如三个人抽签, 而只有一个人能够抽中, 因此每个人抽中的机会是1/3. 假定用A_1,A_2,A_3分别代表这三个人抽中的事件, 那么, $P(A_1)=P(A_2)=P(A_3)=1/3$. 但是由于一个人抽中, 其他人就不可能抽中, 所以, 这三个事件不独立. 刚才的乘法规则不成立, 这时, $P(A_1\cap A_3)=P(A_1\cap A_2)=P(A_2\cap A_3)=0$, 而如错用乘法规则应该是$(1/3)^2=1/9$. 但是可以计算条件概率, 比如第一个人抽到(事件A_1), 则在这个条件下其他两个人抽到的概率都为0, 记为$P(A_2|A_1)=P(A_3|A_1)=0$. 如果第一个人没有抽到(事件A_1^C), 那么其他两人抽到的概率均为1/2, 记为$P(A_2|A_1^C)=P(A_3|A_1^C)=1/2$. 一般地, 在一个事件$B$已经发生的情况下, 事件$A$发生的条件概率定义为(贝叶斯公式)

$$P(A|B)=\frac{P(A\cap B)}{P(B)},\quad P(B)\neq 0.$$

思考一下:

1. 在得奖品的抽签中, 先抽和后抽是否机会是一样的? 为什么?
2. 一个有10个主要部件的机械中, 如果第i个出问题的概率为p_i, 而且它们是否出问题互相独立, 那么全都不出问题的概率是多少?

4.3 变量的分布

随机变量取一切可能值或范围的概率或概率的规律称为**概率分布(probability distribution, 简称分布)**. 有一些概率分布可以用表或各种图来表示, 一些可以用公式来表示, 当然, 很多分布很难表示出来. 一个概率分布是和某**总体(population)**也称为**样本空间(sampling space)**相联系的. 在第二章中我们提到了抽样调查时的总体(或有限总体), 那是没有和任何概率相联系的实际存在. 我们在第二章的注中也提到了在确定了抽样方法之后, 这个有限总体就可能与概率有关的总体有某种联系了, 并且可能对诸如总体比例等进行推断. 这里的总体或样本空间为一个抽象的空间, 它是由某种试验的所有可能结果点组成的, 这些结果的获得都服从某种概率规律. 因此, 一个总体(样本空间)是由一个取值范围及相连的概率所组成的. 因此给出了概率分布就等于知道了总体. 一些用数学语言表示的概率分布有一些理论参数, 称为**总体参数(population parameter)**. 在第三章介绍过基于样本数据的样本均值、样本标准差和样本方差等概念. 这些样本特征可能是相应的总体特征的反映. 我们也有描述变量"位置"的总体均值、总体中位数、总体百分位数以及描述变量分散(集中)程度的总体标准差和总体方差等概念. 具体公式见本章后面小结.

4.3.1 离散随机变量的分布

离散随机变量只取离散的值[1], 比如骰子的点数、次品的个数、得病的人数等等. 每一种取值都有某种概率. 各种取值点的概率总和应该是1. 当然离散变量不仅限于取非负整数值. 一般来说, 某离散随机变量的每一个可能取值x_i都相应于取该值的概率$p(x_i)$, 这些概率应该同时满足关系

$$\sum_i p(x_i) = 1, \quad p(x_i) \geq 0.$$

满足这样的关系的那些$p(x_i)$就称为该离散随机变量的概率分布. 离散变量取值的个数不一定是有穷的. 例如后面要介绍的Poisson分布的取值范围就是所有的非负整数, 因此有无穷多的可能值.

1. 二项分布

最简单的离散分布应该是抛硬币所基于的概率分布. 比如用p代表得到硬币正

[1]离散变量只在有穷的或者可数的集合中取值. 所谓**可数(countable)**, 意味着集合中每个数目都可以和自然数一一对应(可以用自然数编号).

面的概率, 那么$1-p$则是得到反面的概率. 如果知道$p=1/2$, 就有$1-p=1/2$, 于是这个抛硬币的试验的概率分布也就都知道了. 这种试验可以重复很多次.

这种有两种可能结果的试验有两个特点: 一是各次试验互相独立, 二是每次试验得到一种结果的概率不变(这里是得到正面的概率总是p). 类似于抛硬币的仅有两种结果的重复独立试验被称为**Bernoulli试验(Bernoulli trials)**. 下面的试验都可以近似地看成为Bernoulli试验: 每一个进入某商场的顾客都有购买或不购买商品的两种可能、每个被调查的人士会支持或不支持某种观点、每一个产妇有生出男婴和女婴两种可能等等. 根据这种简单试验的分布, 可以得到基于这个试验的更加复杂事件的概率.

为了叙述方便, 人们通常把Bernoulli试验的两种结果称为"成功"和"失败". 和Bernoulli试验相关的最常见的问题是: 如果进行n次Bernoulli试验, 每次成功的概率为p, 那么成功k次的概率是多少? 这个概率的分布就是所谓的**二项分布(binomial distribution)**. 之所以取这个名字是因为该分布和二项式展开的系数有关(参见本章后面的公式). 这个分布有两个参数, 一个是试验次数n, 另一个是每次试验成功的概率p. 基于此, 二项分布用符号$B(n,p)$或$Bin(n,p)$表示. 由于n和p可以根据实际情况取各种不同的值, 因此二项分布是一族分布, 族内的分布以这两个参数来区分. 根据公式容易得到二项分布$B(n,p)$的(总体)均值为np, 方差为$np(1-p)$, 标准差为$\sqrt{np(1-p)}$. 显然, 一次Bernoulli试验成功与否的概率分布为二项分布的特例$B(1,p)$.

二项分布的概率过去常用二项分布表来查出. 现在从任何统计软件都可以很容易得到这个概率. 在目前统计软件发达的情况下, 对于较复杂的问题所涉及的二项分布一般都自动处理了. 在处理实际问题中很少会遇到直接按照公式计算二项分布概率的情况, 但这里还是给出其一般公式. 下面$p(k)$代表在n次Bernoulli试验中成功k次的概率, p为每次试验成功的概率. 有

$$p(k)=\binom{n}{k}p^k(1-p)^{n-k},\quad k=0,1,...,n$$

这里

$$\binom{n}{k}=\frac{n!}{k!(n-k)!}$$

为二项式系数, 按照不同习惯, 也用C_n^k, C_k^n, $C_{n,k}$, ${}_nC_k$等不同符号表示.

图4.1为用R产生的$B(5,p)$在$p=0.1$到0.9的9个二项分布的条形图, 横坐标是成功次数, 而纵坐标为某个成功次数值上的概率.

从图4.1可以看出, 只有当成功概率等于失败概率时($p=0.5$), 这个分布是对称的, 即在五次试验中成功0次(失败5次)和成功5次(失败0次)的概率都是0.03125, 成功1次(失败4次)的概率和失败1次(成功4次)的概率都是0.15625等等. 而当p为其他不等于0.5的值时, 分布就不对称了.

下面两个表分别为$B(5,0.5)$和$B(5,0.7)$的分布, 它们是用下面R语句计算的:

```
dbinom(0:5,5,.5);dbinom(0:5,5,.7)
```

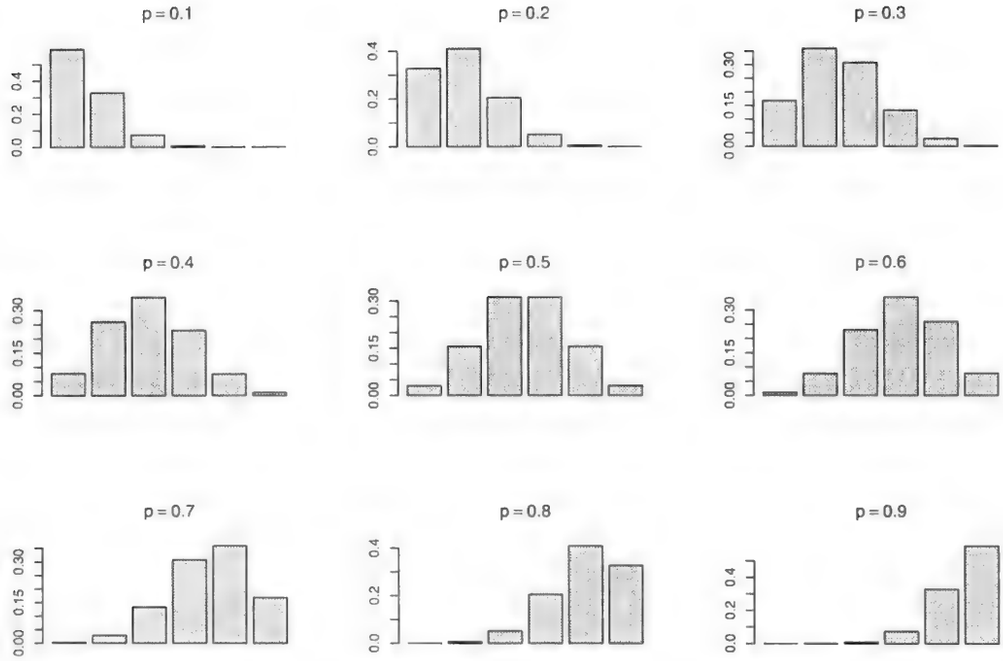

图 4.1 九个二项分布$B(5,p)$ ($p=0.1$到0.9)的概率分布图.

$B(5,0.5)$的分布

成功次数k	0	1	2	3	4	5
概率$p(k)$	0.03125	0.15625	0.3125	0.3125	0.15625	0.03125

$B(5,0.7)$的分布

成功次数k	0	1	2	3	4	5
概率$p(k)$	0.00243	0.02835	0.1323	0.3087	0.36015	0.16807

2. 多项分布

和二项分布最类似的分布是二项分布的推广, 称为**多项分布(multinomial distribution)**. 二项分布的每次试验中只有两种可能的结果, 而多项分布则在每次试验中有多种可能的结果. 比如在调查顾客对5个品牌的饮料的选择中, 每种品牌都会以一定的概率中选, 假定这些概率为p_1,p_2,p_3,p_4,p_5. 每次试验的结果只可能有一个, 因此这些概率的和为1, 即$p_1+p_2+p_3+p_4+p_5=1$. 在二项分布中, 人们关心的是在n次试验中成功k次的概率(有了成功k次的概率, 就有了失败$n-k$次的概率). 但是在多项分布问题中(用上面5个品牌的例子说明), 所关心的是在n次试验中(这里是调查), 选择5个品牌的人数分别为m_1,m_2,m_3,m_4,m_5的概率. 自然, $m_1+m_2+m_3+m_4+m_5=n$. 类似于二项分布, 多项分布的符号可以为$M(n;p_1,p_2,p_3,p_4,p_5)$, 也有用"MN"或"$Multi$"来表示的. 当然, 符号并不重要. 一个前面已经谈过多次的多项分布的例子是掷骰子. 这里有六个结果. 如果骰子是公平的, 那么在一次试验中出现每种点数的概率都是1/6. 因此, 在n次掷骰子中, 得到各种点数的数目这就是一个多项分布: $M(n;1/6,1/6,1/6,1/6,1/6,1/6)$.

再比如打靶. 假定每次射击有从0到10环的11种可能, 显然, 一个人在n次类似条件下的射击中, 得到各种环数的数目也可以认为近似地服从一个多项分布.

3. Poisson分布

另一个常用离散分布是**Poisson分布**(翻译成"泊松分布"或"普阿松分布"). 它可以认为是衡量某种事件在一定期间出现的数目的概率. 比如, 在一定时间内顾客的人数、打入电话总机电话的个数、放射性物质放射出来并到达某区域的粒子数等往往被认为近似地服从Poisson分布. 当然, 在不同条件下, 同样事件在单位时间中出现同等数目的概率不尽相同. 比如中午和晚上某商店在10分钟内出现5个顾客的概率就不一定相同. 因此, 和二项分布一样, Poisson分布也是一个分布族. 族中不同成员的区别在于事件出现数目的均值(通常用λ表示)不一样. Poisson分布的可能取值范围为所有非负整数. 参数为λ的Poisson分布变量的概率分布为($p(k)$表示Poisson变量等于k的概率)

$$p(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, 2, ...$$

图4.2为参数为3、6、10的Poisson分布在$k = 0, 1, 2, ..., 20$处的概率图.

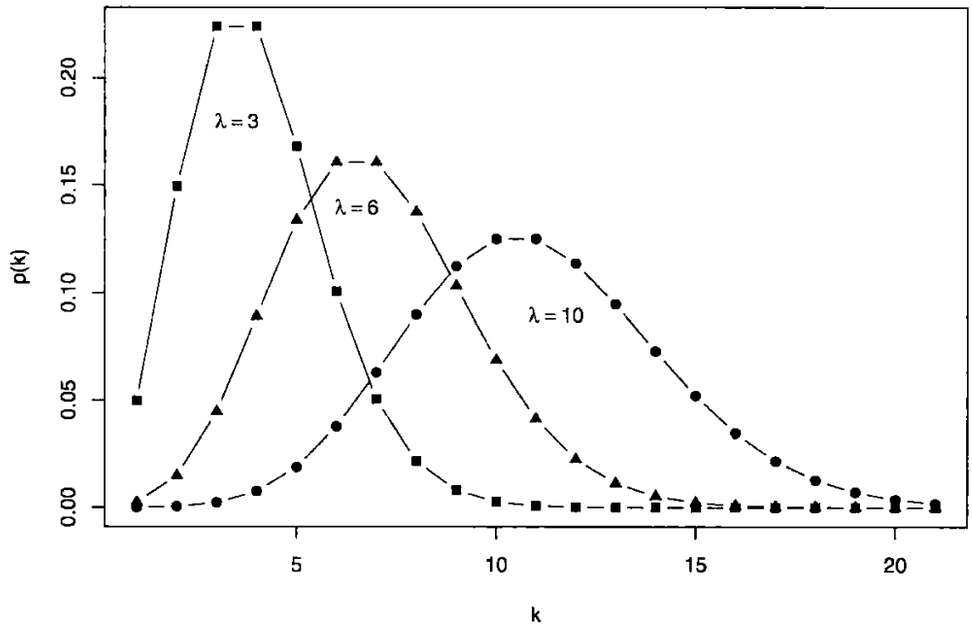

图 4.2　参数λ为3、6、10的Poisson分布(只标出了20之内的部分).

Poisson分布不是对称的, 它在右边有长长的尾巴. 当然, 从图上可以看出尾巴上整数点的概率(纵坐标)非常小. 这里没有用条形图, 而用了散点图的形式. 那些点之间用连线连接主要是为了容易比较这三个分布的形状. 实际上只有在整数点上的概率值才有意义.

Poisson分布的概率过去可以从统计书后面的表中得到, 现在可以从任何统计软件中得到. 不过在目前统计软件发达的情况下, 在处理实际问题中很少会遇到直接用公式计算Poisson分布的概率的情况. 参数为λ的Poisson分布用$P(\lambda)$表示. 由其概率分布可以证明其参数λ既是Poisson变量的总体均值, 也是总体方差, 其标准差为λ的平方根. Poisson分布族中的成员是用不同的λ来区分的.

4. **超几何分布**

超几何分布和有限总体的不放回抽样的实践有密切关系. 假定有一批500个产品, 而其中有5个次品. 质量检查人员随机抽取20个产品进行检查. 如果抽到的20个产品中含有2个或更多不合格产品, 则整个500个产品将会被退回. 这时, 人们想知道, 该批产品被退回的概率是多少? **这种概率就满足超几何分布(hypergeometric distribution)**. 这是一种所谓的"不放回抽样", 也就是说, 一次抽取若干物品, 每检查一个之后并不放回. 这样, 每一个产品都不会被重复检查. 如果是"放回式抽样", 也就是每检查一个就把它放回, 这样再抽取时, 检查过的物品还有可能被抽上, 那么每次抽样时得到次品的概率是一样的, 等于次品的比例, 这就不是超几何分布而是二项分布了. 超几何分布族的成员被三个参数决定: 产品总个数n, 其中不合格产品数目m, 不放回抽样的数目t. 而样本中有x个不合格产品的概率为

$$p(x)=\frac{\binom{m}{x}\binom{n-m}{t-x}}{\binom{n}{t}},\quad x=0,1,...,t.$$

看得出来, 超几何分布和排列组合密切相关. 现在, 计算机软件很容易计算超几何分布, 实际工作者很少有机会自己通过公式用笔和纸计算超几何分布了.

思考一下:

1. 假定有一批500个产品, 而其中有5个次品. 质量检查人员随机抽取20个产品进行检查. 那么放回抽样和不放回抽样时发现2个次品的概率有什么不同? 各自服从什么分布?
2. 在前面离散分布的定义中说概率应该同时满足关系$\sum_i p(x_i)=1$和$p(x_i)\geq 0$, 有人说第二个式子应该为$1\geq p(x_i)\geq 0$, 你觉得有关系吗?
3. 是否可以认为二项分布为多项分布的一个特例?

4.3.2 连续随机变量的分布

许多变量取连续值, 比如高度、长度、重量、时间、距离等等, 它们被称为**连续变量(continuous variable)**. 换言之, 一个随机变量如果能够在一个区间(无论这个区间多么小)内取任何值, 则称之为在此区间内的连续随机变量, 其分布称为**连续型概率分布**. 这时它们的概率分布就很难准确地用描述离散变量概率的

条形图来表示. 让我们想象连续变量观测值的直方图, 如果其纵坐标为相对频数, 那么所有这些矩形条的高度和为1, 而且完全可以重新设置量纲, 使得这些矩形条的面积和为1. 如果不断增加观测值, 并不断增加直方图的矩形条的数目, 这些直方图就会越来越像一条光滑曲线, 其下面的面积和为1. 这种曲线就是所谓**概率密度函数(probability density function, pdf)**, 它简称为密度函数或密度. 图4.3就展示了逐渐增加矩形条数目的直方图和一个形状类似的密度曲线.

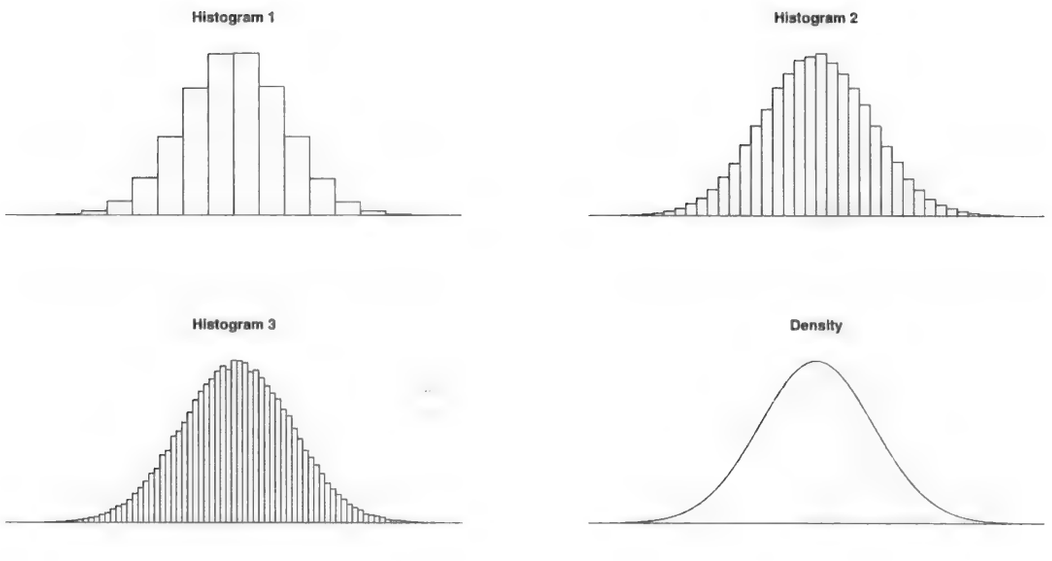

图 4.3　逐渐增加观测值数目和矩形条数目的直方图和一个形状类似的密度曲线.

连续变量落入某个区间的概率就是概率密度函数的曲线在这个区间上所覆盖的面积, 因此, 在理论上, 这个概率就是密度函数在这个区间上的积分. 学过微积分的人都知道, 连续函数在一个点的积分是0(因为曲线下面的面积退化成一条线), 所以, 对于连续变量, 取某个特定值的概率都是零, 而只有变量取值于某个(或若干个)区间的概率才可能大于0. 和离散变量所有取值的概率和为1类似, 连续变量密度函数曲线(这里用f表示)下面覆盖的总面积为1, 即

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

当然, 连续随机变量也有描述变量"位置"的总体均值、总体中位数、总体百分位数以及描述变量分散(集中)程度的总体标准差和总体方差等概念. 具体公式见本章后面小结.

下面介绍几种常见连续变量的分布.

1. 正态分布

在北京市场上的精制盐很多是一千克袋装, 上面标有"净含量1kg"的字样. 但当你用稍微精确一些的天平称那些袋装盐的重量时, 会发现有些可能会重些, 有些可能会轻些, 但都是在1kg左右. 多数离1kg不远, 离1kg越近就越可能出

现, 离1kg越远就越不可能. 一般认为这种重量分布近似地服从最常用的**正态分布(normal distribution**, 又叫**高斯分布, Gaussian distribution**). 近似地服从正态分布的变量很常见, 像测量误差、商品的重量或尺寸、某年龄人群的身高和体重等等. 此外, 在一定条件下, 许多不是正态分布的样本均值在样本量很大时, 也可用正态分布来近似.

正态分布的密度曲线是一个对称的钟型曲线(最高点在均值处). 图4.3所描述的分布就是正态分布. 正态分布也是一族分布, 各种正态分布根据它们的均值和标准差不同而有区别. 因此一个正态分布用$N(\mu,\sigma)$表示, 其中μ为(总体)均值, 而σ为(总体)标准差. 正态分布也常用$N(\mu,\sigma^2)$来表示, 这里σ^2为(总体)方差(标准差的平方). 当然这里的均值和标准差是总体参数, 而不是样本均值和样本标准差. 这些总体参数在实际问题中是不知道的, 但可以估计, 比如用样本均值和样本标准差来估计总体均值和总体标准差.

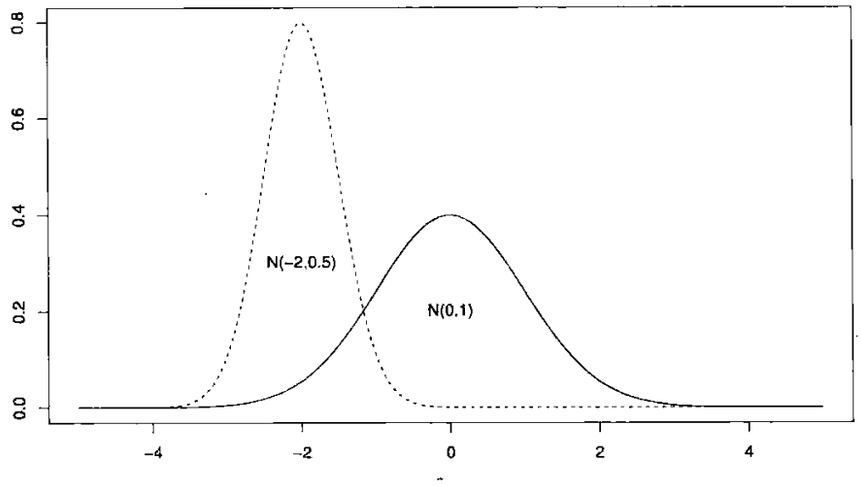

图 4.4 两条正态分布的密度曲线. 左边是$N(-2,0.5)$分布, 右边是$N(0,1)$分布.

图4.4就是放在一张图中的两条正态分布的曲线. 左边的是$N(-2,0.5)$分布, 右边的是$N(0,1)$分布. 均值为0, 标准差为1的正态分布$N(0,1)$称为**标准正态分布(standard normal distribution)**. 标准正态分布的密度函数通常用$\phi(x)$表示. 任何具有正态分布$N(\mu,\sigma)$的随机变量X都可以用简单的变换(减去其均值μ, 再除以标准差σ)$Z=(X-\mu)/\sigma$而成为标准正态随机变量. 这种变换和标准得分的意义类似.

当然, 和所有连续变量一样, 正态变量落在某个区间的概率就等于在这个区间上密度曲线下面的面积. 比如, 标准正态分布变量落在区间(0.51,1.57)中的概率, 就是在标准正态密度曲线下面在0.51和1.57之间的面积. 图4.5表示了这个面积. 利用统计软件的有关函数(后面要介绍)很容易得到这个面积等于0.24682, 也就是说, 标准正态变量在区间(0.51,1.57)中的概率等于0.24682. 记密度函数为$\phi(x)$, 那么

这个面积等于积分.

$$\int_{0.51}^{1.57} \phi(x)dx = 0.24682.$$

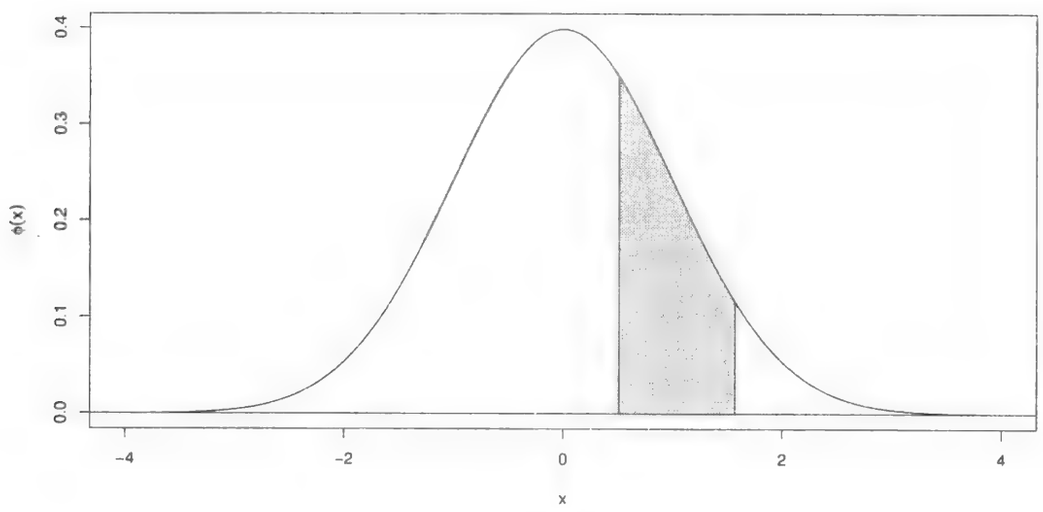

图 4.5　标准正态变量在区间(0.51, 1.57)中的概率(阴影部分面积).

现在引进总体的下侧分位数、上侧分位数以及相应的尾概率的概念. 对于连续型随机变量X, **α下侧分位数**(又称为**α分位数**, α-quantile)定义为满足关系$P(X \leq x_\alpha) = \alpha$的数$x_\alpha$, 这里的$\alpha$称为**下(左)侧尾概率(lower/left tail probability)**. 而**α上侧分位数**(又称**α上分位数**, α-upper quantile)定义为它满足关系$P(X \geq x_\alpha) = \alpha$的数$x_\alpha$, 这里的$\alpha$称为**上(右)侧尾概率(upper/right tail probability)**. 对于一般的分布, 分位数的定义稍微复杂一些[1]. 显然, 对于连续分布, α上侧分位数等于$(1-\alpha)$下侧分位数, 而$(1-\alpha)$上侧分位数等于α下侧分位数. 通常用z_α表示标准正态分布的α上侧分位数, 即对于标准正态分布变量Z, 有$P(Z > z_\alpha) = \alpha$. 图4.6表示了0.05上侧分位数$z_\alpha = z_{0.05}$及相应的尾概率$(\alpha = 0.05)$. 有些书用符号$z_{1-\alpha}$而不是z_α来表示α上侧分位数, 因此在看参考文献时要注意符号的定义.

在统计推断过程中, 往往需要要对正态分布变量进行变换. 这些变换之后的变量, 作为正态分布变量的函数, 就不一定是正态分布了. 只有正态分布变量的线性组合才会仍然是正态分布. 下面介绍由正态分布导出的三种分布. 这些是以后章节中经常会遇到的分布.

2. χ^2分布

由正态变量导出的分布之一是**χ^2分布(chi-square distribution**, 也翻译

[1]对于一般的分布, 总体的α下侧分位数定义为满足$P(X < x_\alpha) \leq \alpha \leq P(X \leq x_\alpha)$的$x_\alpha$, 而$\alpha$上侧分位数定义为满足$P(X > x_\alpha) \leq \alpha \leq P(X \geq x_\alpha)$的$x_\alpha$. 这些分位数一般并不一定唯一, 只有对于连续分布, 分位数才唯一.

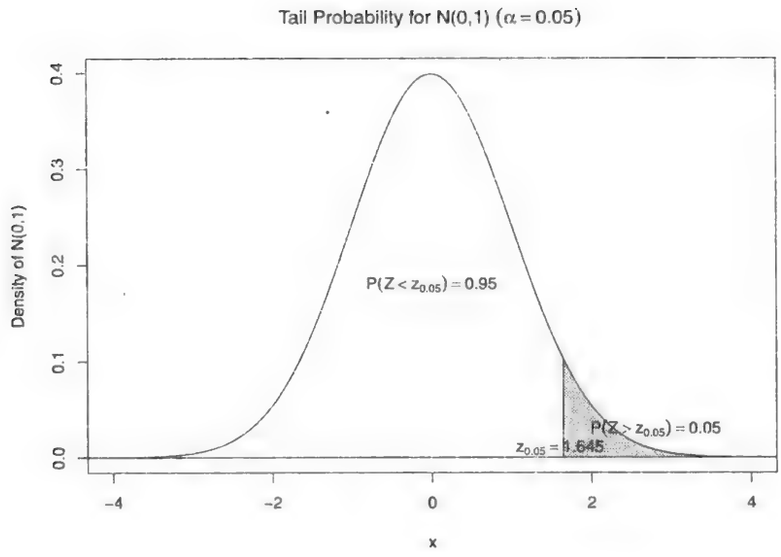

图 4.6　表示$N(0,1)$分布右侧尾概率$P(Z > z_\alpha) = \alpha$的示意图(这里$\alpha = 0.05$).

为**卡方分布**). n个独立标准正态变量的平方和称为有n**个自由度的**χ^2**分布**. 记为$\chi^2(n)$. 更一般地, 若干个独立的χ^2分布变量的和也有χ^2分布, 其自由度等于那些χ^2分布自由度之和. χ^2分布也是一族分布, 由该族成员的不同自由度来区分. χ^2分布在后面要介绍的一些检验中会用到. 由于χ^2分布变量为正态变量的平方和, 它不会取负值.

图4.7为三个不同自由度的χ^2分布密度图. 该分布在一般的统计书中都有概率表. 而在计算机统计软件的解题过程中, 一般都会自动算出所需要的与χ^2分布有关的结果.

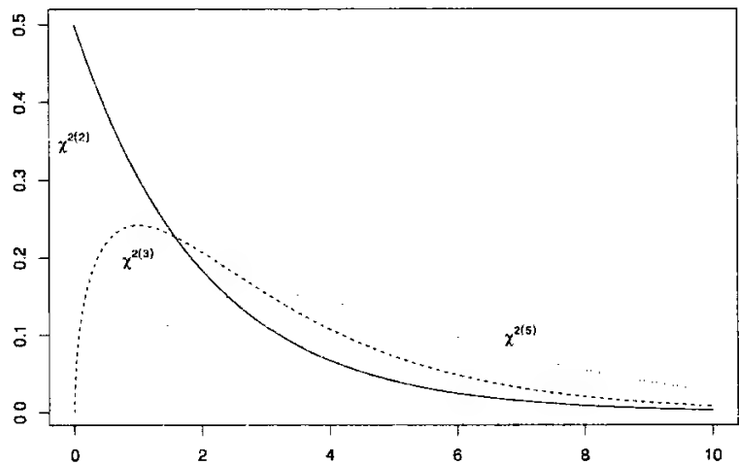

图 4.7　自由度为2、3、5的χ^2分布密度图(只显示了小于10的部分).

3. t分布

正态变量的样本均值也是正态变量. 在统计推断中往往希望利用它减去总体均值再除以均值的总体标准差来得到标准正态变量. 在这个变换中, 如果用均值的样本标准差来代替其未知的总体标准差时, 即用$(\bar{x}-\mu)/(s/\sqrt{n})$代替$(\bar{x}-\mu)/(\sigma/\sqrt{n})$, 得到的结果分布就不再是标准正态分布了. 它的密度曲线看上去有些象标准正态分布, 但是中间瘦一些, 而且尾巴长一些. 这种分布称为**t分布(t-distribution, 或学生分布, Student' s t)**. 之所以叫t分布是因为提出者Gosset用t来表示这个变量, 而发表有关论文时, Gosset用的假名字Student, 因此也叫做学生分布. 不同的样本量通过标准化所产生的t分布也不同, 这样就形成了一族分布. t分布族中的成员是以自由度来区分的. 这里的自由度等于样本量减去1(如果样本量为n, 刚才定义的t分布的自由度为$n-1$, 参见本章后面小结.).

由于产生t分布的情况不只上面一种, 简单说自由度就是样本量减1是不准确的. 自由度这个概念还出现在其他分布之中, 基本上是信息量大小的一个度量. 在t分布中, 如果自由度趋于无穷, 那么t分布就是标准正态分布了. 一个有k个自由度的t分布用$t(k)$表示, 当然也有用$t_{(k)}$或t_k表示的. 图4.8展示了标准正态分布$N(0,1)$和自由度等于1的$t(1)$分布的密度函数曲线. 可以看出t分布两边尾巴比较长. 但是当自由度增加时, 它的分布就逐渐接近标准正态分布了. 因此, 在大样本时, 可以用标准正态分布来近似t分布. t分布还可以用χ^2分布导出, 这将在最后小结中说明.

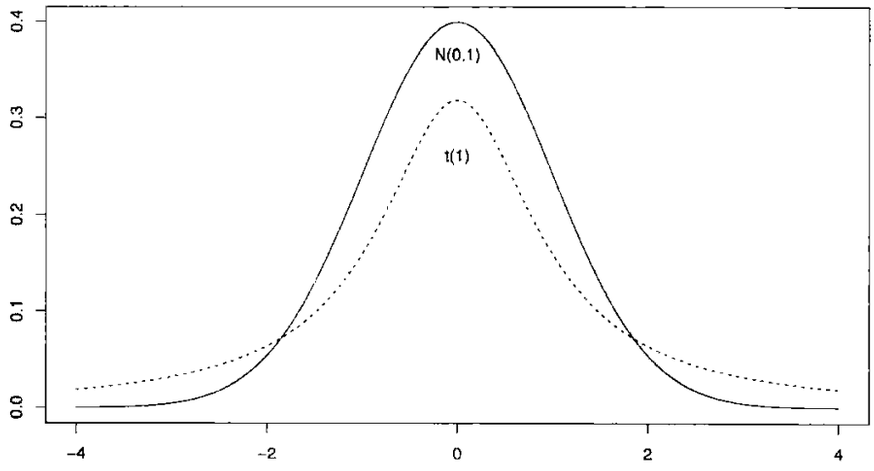

图 4.8　标准正态分布和t(1)分布的密度图.

通常用t_α表示t分布相应于右侧尾概率α的t变量的α上侧分位数, 即对于t分布变量T, 有$P(T>t_\alpha)=\alpha$. 在突出自由度时, 也用$t_{n,\alpha}$, 也有的书用$t_{\alpha-1}$或$t_{n,\alpha-1}$表示. 图4.9表示了自由度为2的$t(2)$分布右边的尾概率($\alpha=0.05$).

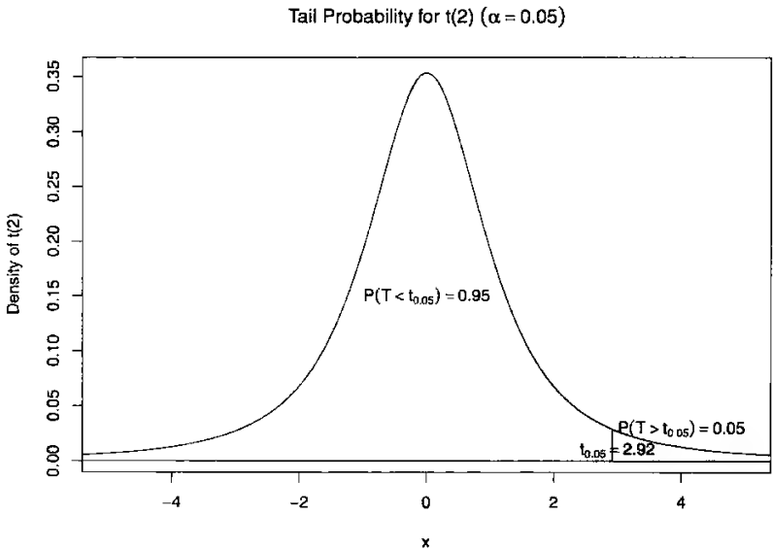

图 4.9 表示t(2)分布右侧尾概率$P(T > t_\alpha) = \alpha$的示意图(这里$\alpha = 0.05$).

4. F分布

两个独立χ^2分布变量(在除以它们各自自由度之后)的比称为**F分布**变量, 而两个χ^2分布的自由度则为F分布的自由度, 因此, F分布有两个自由度: 第一个自由度等于在分子上的χ^2分布的自由度, 第二个自由度等于在分母的χ^2分布的自由度. 人们很少手算F分布概率, 通常都是计算机代劳了. 图4.10为自由度分别为(3,20)和(50,20)的两个F分布密度图. 可以看出, 当第二个自由度相同时, 第一个自由度越小, 峰越靠近左边.

5. 均匀分布

均匀分布(uniform distribution)是最简单的连续型分布. 它的取值范围是一个区间, 比如(a, b). 均匀分布随机变量X取值在该区间的一个子区间的概率等于该子区间宽度与区间(a, b)宽度$b - a$之比. 比如区间(a, b)为(0,1)区间, 那么均匀分布变量X落入$(0.3, 0.7)$的概率为$(0.7 - 0.3)/(1 - 0) = 0.4$. 显然, 均匀分布的密度函数在$(a, b)$区间为常数$1/(b - a)$, 而在该区间外为零. 这种形状为一个矩形, 因此均匀分布也称为**矩形分布(rectangular distribution)**. 图4.11展示了在区间$(0, 1)$上的均匀分布的密度函数.

4.3.3 累积分布函数

在前面离散分布的情况可以用$p(x)$表示该变量取值x的概率, 如果用大写英文字母X表示相应分布的随机变量, 那么概率$P(X = x) = p(x)$. 如果X的取值范围为整数, 则有

$$P(m \le X \le n) = \sum_{k=m}^{n} p(k) = p(m) + p(m+1) + \cdots + p(n)$$

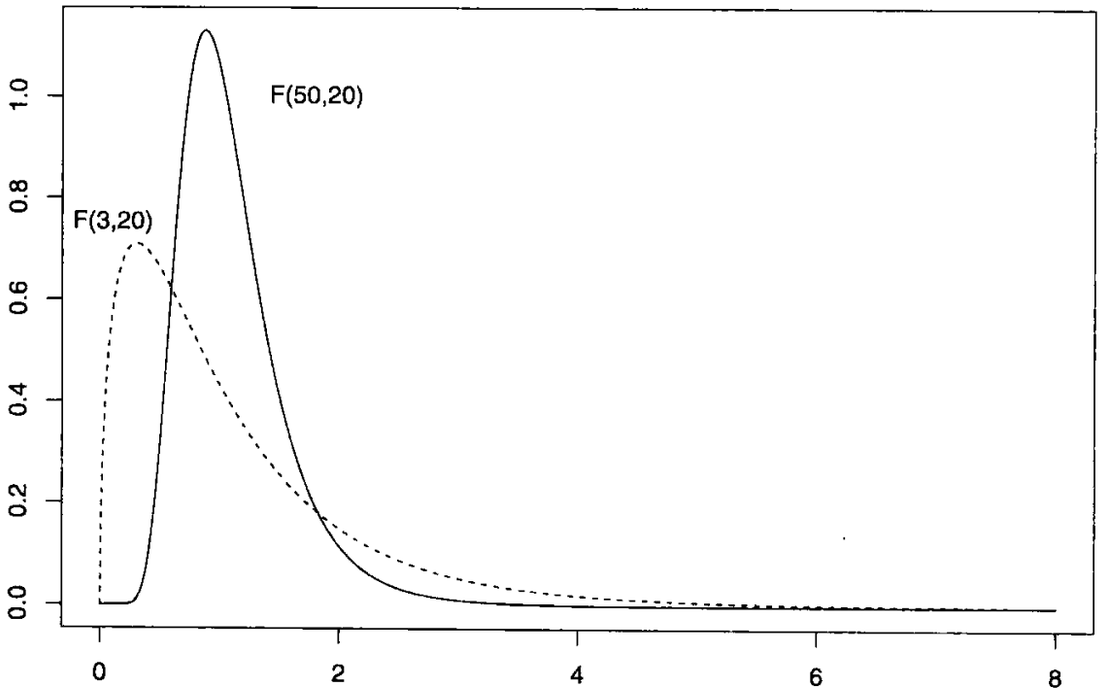

图 4.10 自由度为(3, 20)和(50, 20)的F分布密度曲线图.

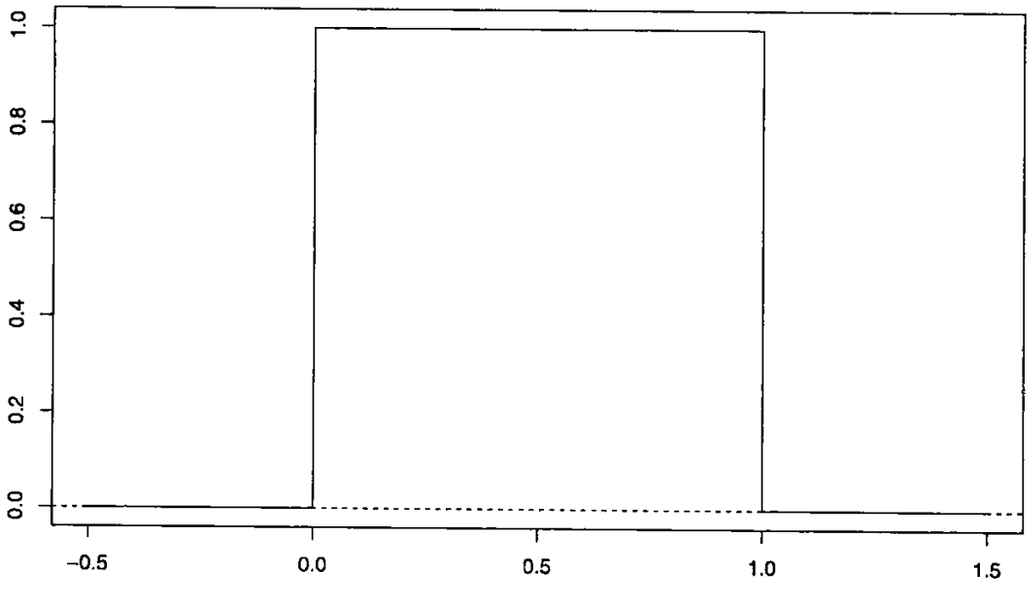

图 4.11 在区间(0,1)上的均匀分布的密度函数.

$$= P(X \leq n) - P(X \leq m-1).$$

在连续分布的情况, 可以用$f(x)$表示密度函数, 则概率(注意在连续分布中, 单独点的概率为0, 因此下式中的不等式中的等号可以去掉)

$$P(a \leq X \leq b) = \int_a^b f(x)dx = P(X \leq b) - P(X \leq a).$$

为了计算概率, 只知道密度函数对于查表或应用软件来得到已知分布的概率是不方便的, 最好能够知道随机变量小于或等于某值的概率. 在上面公式中, 如果知道了下面的值:

$$P(X \leq n),\ P(X \leq m-1),\ P(X \leq b),\ P(X \leq a),$$

那么, 在对离散分布做计算时就不用做那么多的加法, 而在连续分布时就不用做积分了, 都仅仅做一个减法即可. 这种随机变量小于或等于某个数值的概率$P(X \leq x)$ 就称为**累积分布函数(cumulative distribution function**, 简称cdf)或**分布函数**. 累积分布函数概念的引进, 对于查表或使用软件得到概率(根据上面两个公式)是很方便的. 多数概率分布表都是以累积分布函数的形式出现的. 在后面介绍软件时, 还要举例说明如何利用累积分布函数.

思考一下:

1. 讨论离散随机变量和连续随机变量之间的区别.
2. χ^2分布、F分布、t分布都是由正态分布导出的分布, 它们在统计中很常见, 并不是因为它们在实际数据中很常见, 而是因为它们和常见的正态分布之间的关系. 而正态分布则由于下面要介绍的中心极限定理而变得十分重要. 这几个分布主要出现在和正态总体有关的检验中, 这在后面章节会逐渐出现.
3. 连续型分布取某一指定值的概率为零, 但由于四舍五入, 连续型分布实现的记录为离散的. 比如年龄应该是连续变量, 但记录时大多精确到年或月(最多到日). 这样连续型分布的记录值则仅仅取一些离散值了. 这应该解释某些诸如一些连续变量的实现值相同等现象.
4. 注意, 真实世界变量的分布大都是不知道的, 因此人们希望用少数可以用数学语言表示的分布族来近似地描述真实分布. 这种做法有很大的局限性. 非参数统计以及机器学习的方法就是摆脱这种束缚的实践.

4.4 抽样分布、中心极限定理

我们希望利用样本, 特别是通过作为样本函数的样本统计量来了解总体, 来对总体参数进行推断. 这些样本统计量包括前面提到过的样本均值、样本中位数、样本标准差以及由它们组成的函数. 这些样本统计量对于不同的样本(但有相同的样本量)会取不同的值, 也就是说, 具有相同样本量的样本统计量作为随机

样本的函数也是随机的, 也有自己的分布. 这些分布就称为**抽样分布(sampling distribution)**. 为了理解抽样分布的直观意义, 我们来看每次掷5次公平骰子的试验(样本量$n=5$), 来看样本均值的变化. 下表记录了前15次每次得到的结果和该次的样本均值.

试验编号	5次掷骰子的结果					样本均值
	X_1	X_2	X_3	X_4	X_5	$\bar{X}$
1	4	5	6	4	2	4.2
2	1	3	6	6	3	3.8
3	1	1	4	2	6	2.8
4	2	6	5	1	2	3.2
5	6	2	2	3	3	3.2
6	4	1	2	1	5	2.6
7	3	3	4	1	4	3.0
8	4	1	4	5	1	3.0
9	3	6	4	5	6	4.8
10	5	5	6	4	5	5.0
11	1	2	3	3	2	2.2
12	1	1	5	2	2	2.2
13	3	3	4	3	4	3.4
14	5	1	5	5	4	4.0
15	5	6	6	1	4	4.4

显然这些样本均值都和真正的总体均值$(1+2+3+4+5+6)/6=3.5$有些差别. 这15个样本均值的平均(均值)为3.453333, 比表中的哪一次试验的均值都接近总体均值. 这使得人们对这些样本均值的分布感兴趣.

假定一个连续分布的变量X的n个观测值组成一个样本. 如果X的总体均值为μ, 而总体标准差为σ, 这两个总体参数通常是未知的. 现在, 我们主要关注总体的均值μ. 从这个样本, 我们还可以计算样本均值$\bar{X}$和样本标准差s. 当然样本均值可以用来估计μ的值(下一章会介绍). 具体这种估计的好坏, 依赖于样本均值的抽样分布. 样本均值作为随机变量有如下的性质(注意, 这里并没有假定X的分布):

1. 样本均值$\bar{X}$的抽样分布的总体均值等于μ.

2. 样本均值$\bar{X}$的抽样分布的总体标准差等于$\sigma/\sqrt{n}$, 显然, 样本量越大, $\bar{X}$的标准差越小.

3. 即使X的分布不是正态, 那么在很一般的条件下, 当样本量增加时, $\bar{X}$的分布趋近于正态分布$N(\mu,\sigma/\sqrt{n})$. 这就是所谓的**中心极限定理(Central Limit Theorem,** 缩写为CLT)[1].

[1]中心极限定理成立的一个充分条件是, 样本点是独立的, 来自一个总体(同分布), 总体均值存在, 并且有非零有限总体方差.

在上面第二条中把样本均值$\bar{X}$的抽样分布的总体标准差公式中的σ换成样本标准差s, 得到的$s/\sqrt{n}$就是第三章引进的**均值的标准误差(standard error of mean)**. 它是对$\sigma/\sqrt{n}$的一个近似. 中心极限定理是概率论最出色的定理之一. 为了直观地说明它的意义. 我们从在(0,1)的均匀分布对于四种样本量大小$n = 1, 3, 10, 100$分别取600个样本, 在每个样本算出均值. 这样, 对每一种样本量都有600个均值, 用这些均值画直方图(图4.12). 可以看出, 样本量越大, 均值的直方图越像正态变量的直方图, 而且数据的分散程度也越小(越集中).

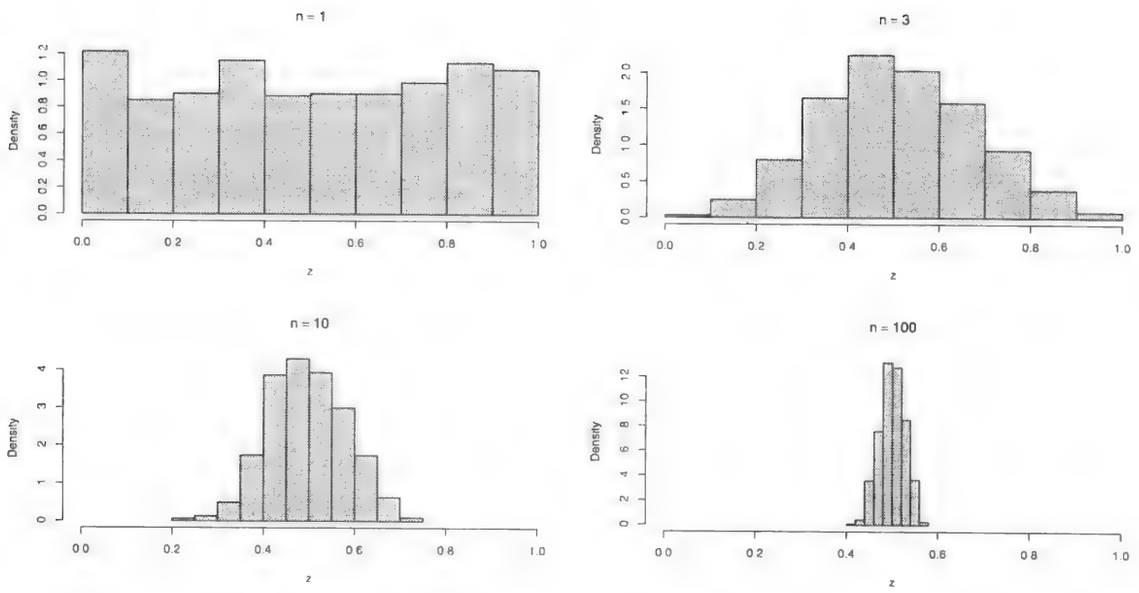

图 4.12 不同样本量的各600个均匀分布样本均值$\bar{x}$的直方图.

统计量的变换

在比较样本均值$\bar{X}$和假定的总体均值μ时, 仅仅考虑差值$\bar{X} - \mu$本身往往不如研究它的某种有同样意义的变换, 以使得容易发现其分布. 类似地, 要比较样本标准差s和假定的总体标准差σ, 也不能仅仅考虑$s^2 - \sigma^2$, 也要进行某种变换, 使得变换过的统计量有某种容易掌握的分布. 一些变换的统计量公式和意义在4.6.2节介绍. 虽然我们不要求记住这些公式, 但对于理解后面关于推断的章节所用的一些统计量不无好处.

思考一下:

1. 抽样分布涉及的是统计量的分布, 我们要用统计量做统计推断就往往需要这些分布. 比如后面马上要讨论的小概率的计算(和后面的假设检验密切相关) 就依赖于这些分布.

2. 本书所涉及的抽样分布主要是前一节的χ^2分布、F分布、t分布. 但要注意, 如果没有正态总体的假定, 或者中心极限定理的应用, 是不会有这些分布的.

4.5 用小概率事件进行判断

判明一个事情的真伪, 需要用事实说话. 在统计中事实总是来源于数据. 下面看小概率事件如何起作用. 假定某药厂声称该厂生产的某种药品有60%的疗效. 但是当实际调查了100名使用该药物的患者之后, 发现最多有40名患者服后有效. 这个数据是否支持药厂的说法呢? 药厂所支持的模型可以看成是一个参数p为0.6的Bernoulli试验模型. 100名患者的服药, 实际上等于进行了100次试验. 这就是二项分布$B(100, 0.6)$模型. 由于使用了药厂的0.6的成功概率. 这个模型是基于药厂的观点的. 我们可以基于这个模型计算100名患者中有少于或等于40名患者治疗有效的概率$P(X \leq 40)$. 通过计算机(R代码`pbinom(40,100,.6)`)或查表, 容易得到该概率为0.000042. 这说明, 如果药厂正确, 那么只有40名患者有效这个事实是个小概率事件, 即少于或等于40名患者有效的可能只有十万分之四多一点. 这样在药厂的观点和事实之间有了矛盾. 是事实准确还是药厂准确呢? 显然人们一般不会认为药厂的说法可以接受. 这样, 就利用小概率事件来拒绝了药厂的说法. 这种用小概率事件对假定的模型进行判断是后面将要介绍的假设检验的基础.

思考一下:

1. 人们在看了上面的例子的数据之后, 可能会觉40名患者服后有效, 就完全可以否定厂方所说的60%有效的说法. 这没有错. 但是, 不经过计算, 无法得出这种概率仅仅约为十万分之四左右的小概率结论. 这也是定量分析所做出的结论比定性结论更强大的原因.

2. 注意, 在上例中, 怀疑的是厂方的说法, 因此计算概率也要以这种说法为基础(二项分布模型是基于厂方60%有效的说法), 发生矛盾, 则说明厂方有问题.

4.6 小结

4.6.1 本章的概括和公式

这一章介绍了概率的概念以及得到概率的途径. 这包括利用等可能事件来得到概率、利用相对频数或频率来近似概率、主观概率等. 还给出了概率的一些运算规则. 最后介绍了一些常用离散和连续变量的分布、中心极限定理及抽样分布. 虽然后面各章都涉及分布和概率的概念, 但由于使用计算机, 直接计算概率的机会并不多. 本章涉及的公式如下.

集合记号: 集合A和B的并记为$A \cup B$, 集合A和B的交记为$A \cap B$, 集合A和B互补记为$B = A^C$或$A = B^C$. 在概率中的事件就相当于集合论中的集合.

互补事件的概率: $P(A) + P(A^C) = 1$, 或者$P(A^C) = 1 - P(A)$, 或者$P(A) = 1 - P(A^C)$.

加法规则:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

特殊情况: 当A和B互斥时$P(A \cap B) = 0$, 所以$P(A \cup B) = P(A) + P(B)$.

乘法规则: 当A和B独立时$P(A \cap B) = P(A)P(B)$.

条件事件的关系(贝叶斯定理):

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

或

$$P(B|A) = \frac{P(A \cap B)}{P(A)};\ P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

如果A和B独立, 则$P(B) = P(B|A)$及$P(A) = P(A|B)$. 反之亦然.

随机变量取一切可能值或范围的概率称为**概率分布(probability distribution)**. 一个离散变量X的概率分布由其可能取值x_k的概率$p(x_k) = P(X = x_k)$来描述.

而连续变量的概率分布由其取值范围内的点x的**概率密度函数(probability density function, pdf)** $f(x)$来描述. 连续随机变量在单独点上的概率为零, 但可以利用积分得到在某区间上的概率. 如果用X表示该连续随机变量, 那么X在区间(a, b)上的概率为

$$P(a < X < b) = \int_a^b f(x)dx.$$

另外还有**累积分布函数(cumulative distribution function, cdf)**的概念, 简称为**分布函数**. 它是随机变量小于或等于某数x的概率, 记为$F(x)$. 对于具有分布$p(x_k)$的离散变量, 分布函数

$$F(x) = P(X \leq x) = \sum_{x_k \leq x} p(x_k).$$

而对于具有分布密度$f(x)$的连续变量, 分布函数

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx.$$

对应于样本均值和样本标准差等特征的是相应的总体特征. 类似地, 总体均值(又称为数学期望)是描述总体的位置参数, 而总体方差和标准差描述总体分布的分散程度.

概率分布为$p(x_k)$的离散变量X的**总体均值**(又称为X的**数学期望**, 记为$\boldsymbol{E}(X)$)定义为

$$\mu = E(X) = \sum_k x_k p(x_k).$$

这和样本均值的定义类似, 只不过权函数不是$1/n$, 而是相应点的概率. 而该变量

的**总体方差**(记为$\boldsymbol{V}ar(X)$)定义为

$$\sigma^2 = Var(X) = \sum_k (x_k - \mu)^2 p(x_k),$$

而总体标准差是方差的平方根

$$\sigma = \sqrt{Var(X)} = \sqrt{\sum_k (x_k - \mu)^2 p(x_k)}.$$

概率密度函数为$f(x)$的连续变量X的**总体均值**(又称为X的**数学期望**, 记为$\boldsymbol{E}(X)$)定义为

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx.$$

而该变量的**总体方差**(记为$\boldsymbol{V}ar(X)$)定义为

$$\sigma^2 = Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx,$$

而总体标准差是方差的平方根

$$\sigma = \sqrt{Var(X)} = \sqrt{\int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx}.$$

总体的**α下侧分位数**(又称为**α分位数**)定义为满足$P(X < x_\alpha) \leq \alpha \leq P(X \leq x_\alpha)$的$x_\alpha$, 而**$\alpha$上侧分位数**定义为满足$P(X > x_\alpha) \leq \alpha \leq P(X \geq x_\alpha)$的$x_\alpha$. 这些分位数一般并不一定唯一, 只有对于连续分布, 分位数才唯一. 对于连续型随机变量X, α下侧分位数为满足关系$P(X \leq x_\alpha) = \alpha$的数$x_\alpha$, 这里的$\alpha$称为**下(左)侧尾概率**, 而**$\alpha$上侧分位数**(又称**$\alpha$上分位数**, **上$\alpha$分位数**) 定义为满足关系$P(X \geq x_\alpha) = \alpha$的数$x_\alpha$, 这里的$\alpha$称为**上(右)侧尾概率**.

随机变量X的**总体中位数**定义为满足$P(X < m) \leq 0.5 \leq P(X \leq m)$的$m$. 随机变量$X$的**总体$k$百分位数**定义为满足$P(X < q) \leq k\% \leq P(X \leq q)$的$q$. 如果令$\alpha = k\%$, 则这个定义也可以说成随机变量$X$的**总体$\alpha$分位数**为满足$P(X < q) \leq \alpha \leq P(X \leq q)$的$q$. 显然, 作为分位数特例的总体中位数为50百分位数或0.5分位数.

在具体分布方面, 我们首先介绍了一些离散变量的分布. 其中包括基于一系列独立可重复的Bernoulli试验的二项分布、描述有多个可能试验结果的多项分布、描述一些事件发生次数的Poisson分布以及涉及不放回抽样的超几何分布. 这些分布都可以利用公式、表格或软件计算. 下面是它们的公式:

总体均值和总体方差的一些性质: 对于均值, 如果$E(X) = \mu$, 则对于任何常数a和b, $E(aX + b) = aE(X) + b = a\mu + b$. 此外, 对于两个随机变量$X$和$Y$, 有$E(aX + bY) = aE(X) + bE(Y)$. 如果$X$和$Y$独立, 则$E(XY) = E(X)E(Y)$. 对于方差, 如果$Var(X) = \sigma^2$, 那么, $Var(aX + b) = a^2 Var(X) = a^2\sigma^2$. 如果$X$和$Y$独立, 则两个变量和的方差满足$Var(X + Y) = Var(X) + Var(Y)$. 作为例子, 如果$X_1, ..., X_n$皆为来自均值为$\mu$, 方差为$\sigma^2$的独立观测值组成的样本, 那

么样本均值$\bar{X}=\sum_{i=1}^{n}X_i/n$的均值$E(\bar{X})$还是$\mu$, 而方差$Var(\bar{X})$为$\sigma^2/n$, 标准差为$\sigma/\sqrt{n}$. 显然, 常数的方差等于0.

二项分布$B(n,p)$: 下面$p(k)$代表在n次Bernoulli试验中成功的次数的概率, p为每次试验成功的概率. 有

$$p(k)=\binom{n}{k}p^k(1-p)^{n-k},\quad k=0,1,...,n$$

这里

$$\binom{n}{k}=\frac{n!}{k!(n-k)!}$$

为二项式系数, 按照不同习惯, 也用C_n^k, C_k^n, $C_{n,k}$, ${}_nC_k$等不同符号表示.

多项分布$M(n;p_1,...,p_k)$: 用$p(m_1,...,m_k)$代表多项分布k种可能在n次试验中分别出现$m_1,...,m_k$次的概率, 而$p_1,...,p_k$为一次试验时各种可能出现的概率. 有

$$p(m_1,...,m_k)=\binom{n}{m_1,...,m_k}p_1^{m_1}p_2^{m_2}\cdots p_k^{m_k},\ \sum_{i=1}^{k}m_i=n,\ \sum_{i=1}^{k}p_i=1,$$

这里

$$\binom{n}{m_1,...,m_k}=\frac{n!}{m_1!\cdots m_k!}$$

为多项式系数.

Poisson分布$P(\lambda)$: 参数为λ的Poisson分布变量的概率分布为($p(k)$表示Poisson变量等于k的概率)

$$P(k)=e^{-\lambda}\frac{\lambda^k}{k!},\ k=0,1,2....$$

超几何分布: 在一批n个产品中, 如果有m个不合格产品(即有$n-m$个合格产品), 那么在不放回抽取t个产品中有x个不合格产品的概率为

$$p(x)=\frac{\binom{m}{x}\binom{n-m}{t-x}}{\binom{n}{t}},\ x=0,1,...,t.$$

本章还介绍了一些连续分布. 其中包括最常用的正态分布、χ^2分布、t分布和F分布. 其中后面三种是从正态分布导出的.

正态分布$N(\mu,\sigma)$的密度函数为(在计算机时代, 真正用这个公式来计算正态变量概率的人已经不多了, 这里介绍它是因为它太著名了):

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

标准正态分布$N(0,1)$为均值为0($\mu=0$), 标准差为1($\sigma=1$)的正态分布, 习惯上, 标准正态分布的密度函数和累积分布函数分别用ϕ和Φ表示.

χ^2分布: 如果$X_1, ..., X_n$互相独立, 而且都是标准正态分布$N(0,1)$, 则

$$\sum_{i=1}^{n} X_i^2$$

有自由度为n的χ^2分布, 记为$\chi^2(n)$.

t分布: 假定有一个来自正态分布$N(\mu,\sigma)$的样本, 样本标准差为s, 样本均值为$\bar{X}$, 样本量为n, 那么

$$t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$$

具有自由度为$(n-1)$的t分布$t(n-1)$. 另外一种定义为: 如果X是$N(0,1)$变量, Y是$\chi^2(n)$变量, 而且X和Y独立, 那么

$$t = \frac{X}{\sqrt{Y}}\sqrt{n}$$

为有n个自由度的t分布$t(n)$.

F分布: 如果X是$\chi^2(m)$变量, Y是$\chi^2(n)$变量, 而且X和Y独立, 那么

$$F = \frac{X/m}{Y/n}$$

为具有自由度(m,n)的F分布, 记为$F(m,n)$.

均匀分布: 如果X是在(a,b)区间上的均匀函数, 那么它的分布密度函数为

$$f(x) = \begin{cases} 1/(b-a) & a \le x \le b, \\ 0 & \text{其他地方} \end{cases}$$

统计量的常用变换

下表给出了几种变换, 这些变换的表达式在后面关于推断的章节可能会出现. 表中各行的符号和假定的意义如下(序号为行号):

1. 对于一个正态变量X: 假定其总体均值为μ, 总体标准差为假定的σ. 它的样本量为n的样本均值为$\bar{X}$.

2. 对于一个正态变量X: 假定其总体均值为μ, 总体标准差σ未知. 它的样本量为n的样本均值为$\bar{X}$, 样本标准差为s.

3. 对于一个正态变量X: 总体标准差为假定的σ. 它的样本量为n的样本标准差为s.

4. 对于两个独立正态变量X_1和X_2: 假定其总体均值分别为μ_1和μ_2, 而总体标准差分别为假定的σ_1和σ_2. 它们样本量分别为n_1和n_2的样本均值分别为$\bar{X}_1$和$\bar{X}_2$.

5. 对于两个独立正态变量X_1和X_2: 假定其总体均值分别为μ_1和μ_2, 而总体标准差σ_1和σ_2假定相等. 它们样本量分别为n_1和n_2的样本均值分别为$\bar{X}_1$和$\bar{X}_2$, 样本标准差分别为s_1和s_2.

6. 对于两个独立正态变量X_1和X_2: 假定其总体均值分别为μ_1和μ_2, 而总体标准差σ_1和σ_2假定不相等. 它们样本量分别为n_1和n_2的样本均值分别为$\bar{X}_1$和$\bar{X}_2$, 样本标准差分别为s_1和s_2.

7. 对于两个独立正态变量X_1和X_2: 总体标准差为σ_1和σ_2, 样本标准差分别为s_1和s_2.

8. 对于一个二项分布变量X: n为样本量, π表示假定的总体概率. $p = x/n$为样本比例.

9. 对于两个二项分布变量X_1和X_2: 样本量分别为n_1和n_2, 总体概率分别为π_1和π_2. $p_1 = x_1/n_1$和$p_2 = x_2/n_2$分别为样本比例.

变换的统计量	统计量的分布	性质
$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$	$N(0,1)$	如果$\bar{X}$ 接近μ, 则Z接近0.
$t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$	$t(n-1)$	如果$\bar{X}$ 接近μ, 则t接近0.
$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$	$\chi^2(n-1)$	如果s^2接近σ^2, 则χ^2接近$n-1$.
$Z = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}$	$N(0,1)$	如$\bar{X}_1-\bar{X}_2$接近$\mu_1-\mu_2$, 则Z接近0.
$t = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{s_p^2(1/n_1+1/n_2)}}$	$t(n_1+n_2-2)$	如$\bar{X}_1-\bar{X}_2$接近$\mu_1-\mu_2$, 则t接近0.
$t^* = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$	近似$t(\ell)$	如$\bar{X}_1-\bar{X}_2$接近$\mu_1-\mu_2$, 则t^*接近0.
$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2}$	$F(n_1-1,n_2-1)$	当s_1^2 接近σ_1^2, s_2^2 接近σ_2^2 时, F接近1.
$Z = \frac{p-\pi}{\sqrt{\pi(1-\pi)/n}}$	n_1和n_2大时Z近似$N(0,1)$	当p接近π时, Z接近0.
$Z = \frac{(p_1-p_2)-(\pi_1-\pi_2)}{\sqrt{\pi_1(1-\pi_1)/n_1+\pi_2(1-\pi_2)/n_2}}$	n大时Z近似$N(0,1)$	当p_1接近π_1, p_2接近π_2时, Z接近0.

其中$s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-1}$, $\ell = \left[\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right] \Big/ \left[\frac{s_1^2}{n_1^2(n_1-1)}+\frac{s_2^2}{n_2^2(n_2-1)}\right]$.

4.6.2 本章例题和R语句说明

本章基本没有统计方法的计算课题, 但是涉及一些具体分布的概率计算. 在实际应用中, 这些概率通常都在运用程序包的过程中自动计算. 但是在做习题和理解一些基本概念方面, 需要通过查表或软件来计算. 下面通过例子简要说明如何通过累积分布函数(第4.3.3节)得到这些概率.

例4.1 求正态分布$N(3,1.5)$变量X在区间(2,4)中的概率$P(2<X<4)$. 从4.5节知道(注意: 连续分布概率表示中的不等号中的等号可有可无, 例如, $P(2<X<4) = P(2 \le X \le 4)$)

$$P(2<X<4) = P(X<4) - P(X<2)$$

所以只要知道$P(X<4)$和$P(X<2)$, 就可以很容易得到$P(2<X<4)$. 在R中, 用语句`pnorm(4,3,1.5)-pnorm(2,3,1.5)`立得结果0.4950149. 图4.13表

现了$N(3,1.5)$分布密度曲线下面在区间(2,4)处的面积等于密度曲线在区间$(-\infty,4)$处曲线下面的面积$P(X<4)$和在区间$(-\infty,2)$ 处下面的面积$P(X<2)$之差.

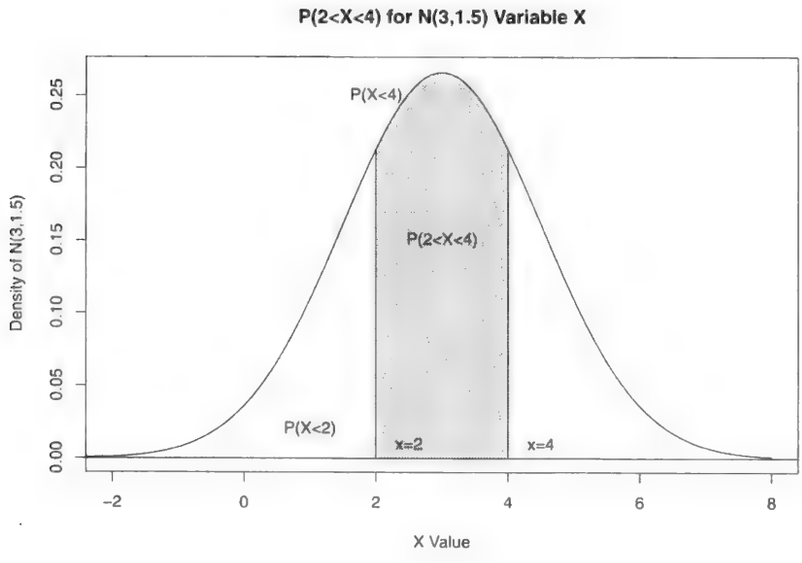

图 4.13　正态$N(3,1.5)$变量X的概率$P(2<X<4)=P(X<4)-P(X<2)$示意图.

注: 在传统统计书中不可能有正态分布族的所有成员的累积分布函数表, 而仅仅有标准正态分布表. 为了使用标准正态分布表, 先把变量按照公式

$$Z=\frac{X-\mu}{\sigma}$$

变换成标准正态变量, 这里Z通常表示标准正态变量, μ为X变量分布的均值, σ为X变量分布的标准差. 然后把公式做相应的改动:

$$\begin{aligned}P(2<X<4) &= P(X<4)-P(X<2)\\ &= P\left(\frac{X-\mu}{\sigma}<\frac{4-\mu}{\sigma}\right)-P\left(\frac{X-\mu}{\sigma}<\frac{2-\mu}{\sigma}\right)\\ &= P\left(Z<\frac{4-3}{1.5}\right)-P\left(Z<\frac{2-3}{1.5}\right)\\ &= P(Z<0.6666667)-P(Z<-0.6666667)\\ &= \Phi(0.6666667)-\Phi(-0.6666667)\\ &= 0.7475075-0.2524925=0.4950149\end{aligned}$$

这里符号Φ专门用来表示标准正态累积分布函数, 它的部分值可以在传统的统计教科书的附表中查到. 可以看出, 用统计软件比查分布表来计算要简单得多. 事实上, 上面分步的计算结果也是用R软件算出来的, 查表不可能得到这么多位有效数字. 注意: 本书不刻意对计算机输出的数字做四舍五入.

例4.2 假定有80%的人喜欢某项产品, 如果随机访问7个人, 则至少3个人喜欢

该产品的概率是多少? 这实际上是求$B(7,0.80)$二项分布变量X等于3到7的概率, 即X大于或等于3的概率(注意, 和连续变量不同, 对于离散随机变量, 不等式中等于号绝不能可有可无!)

$$P(X \geq 3) = P(3 \leq X \leq 7) = \sum_{k=3}^{7} p(k) = p(3) + p(4) + p(5) + p(6) + p(7)$$
$$= P(X \leq 7) - P(X \leq 2) = 1 - P(X \leq 2).$$

注意$P(X \leq 7) = 1$. 因此必须找到$P(X \leq 2)$.

使用R语句`pbinom(2,7,.8,low=F)`或`pbinom(7,7,.8)-pbinom(2,7,.8)`立刻得到结果0.995328.

类似地如果问题是: "随机访问7人, 最多3个人喜欢该产品的概率是多少? "这等于求$P(X \leq 3)$. 在R中, 用语句`pbinom(3,7,.8)`立得结果0.033344.

例4.3 对于尾概率$\alpha = 0.025$, 求标准正态z_α和t(3)分布的t_α. 这当然可以从分布表查到. 但在R中, 用语句`qnorm(0.025,low=F)`和`qt(0.025,3,low=F)`, 立得$z_{0.025} = 1.959964$及$t_{0.025} = 3.182446$. 这属于R中的众多逆函数之一.

4.6.3 生成本章图形的R代码

图4.1是用下面的R代码绘出:

```
par(mfrow=c(3,3))
for(i in seq(.1,.9,.1)){barplot(dbinom(0:5,5,i))
title(main=(substitute(p == that, list(that = i))))}
```

图4.2是用下面代码生成的:

```
plot(dpois(0:20,3),type="b",pch=15,xlab="k",ylab="p(k)")
points(dpois(0:20,6),type="b",pch=17)
points(dpois(0:20,10),type="b",pch=19)
text(c(3.5,6.5,11.5),c(.18,.14,.09),c(expression(lambda==3),
expression(lambda==6),expression(lambda==10)))
```

图4.3是用下面R代码实现的:

```
x=rnorm(100000)
par(mfrow=c(2,2))
hist(x,14,col="blue",axes =FALSE,xlab="",ylab="",
    main="Histogram 1")
hist(x,50,col="blue",axes =FALSE,xlab="",ylab="",
    main="Histogram 2")
hist(x,100,col="blue",axes =FALSE,xlab="",ylab="",
    main="Histogram 3")
z=seq(-4,4,l=1000)
```

```
plot(z,dnorm(z),type="l",axes = FALSE,xlab="",ylab="",
    main="Density")
polygon(c(z[z>-4]),c(dnorm(c(z[z>-4]))),col="blue")
```

图4.4是用下面R代码实现的:

```
x=seq(-5,5,.001)
plot(x,dnorm(x,-2,.5),type="l",lty=2,xlab="",ylab="")
lines(x,dnorm(x)); text(c(-2,0),c(.3,.2),c("N(-2,0.5)","N(0,1)"))
```

图4.5是用下面代码实现的:

```
x=c(seq(-4,4,length=1000))
r1=0.51;r2=1.57;x2=c(r1,r1,x[x<r2&x>r1],r2,r2)
y2=c(0,dnorm(c(r1,x[x<r2&x>r1],r2)),0)
plot(x,dnorm(x),type="l",ylab=expression(phi(x)))
abline(h=0);polygon(x2,y2,col="grey")
```

图4.6是由下面R代码实现的:

```
x=seq(-4,4,length=1000);y=dnorm(x)
plot(x,y,type="l", ylab="Density of N(0,1)");abline(0,0)
r=1.645;polygon(c(r,r,x[x>r]),c(0,dnorm(c(r,x[x>r]))),col="grey")
text(c(0,2.8,1.6),c(.18,.03,.01),c(expression(P(Z<z[0.05])==0.95),
expression(P(Z>z[0.05])==0.05), expression(z[0.05]==1.645)))
title(expression(paste("Tail Probability for N(0,1) ",
    (alpha==0.05))))
```

图4.7是由下面R代码实现的:

```
x=seq(0,10,l=1000);y1=dchisq(x,2);y2=dchisq(x,3);y3=dchisq(x,5);
plot(x,y1,type="l",xlab="",ylab="")
lines(x,y2,lty=2);lines(x,y3,lty=3)
text(c(0,1,7),c(.35,.2,.1),c(expression(chi^2(2)),
expression(chi^2(3)),expression(chi^2(5))))
```

图4.8是由下面R代码实现的:

```
x=seq(-4,4,l=1000);y1=dnorm(x);y2=dt(x,1)
plot(x,y1,type="l",xlab="",ylab="")
lines(x,y2,lty=2);text(c(0,0),c(.2,.35),c("N(0,1)","t(1)"))
```

图4.9是由下面R代码实现的:

```
x=seq(-6,6,length=1000);y=dt(x,2)
r1=2.92;r2=6;x2=c(r1,r1,x[x<r2&x>r1],r2,r2)
y2=c(0,dt(c(r1,x[x<r2&x>r1],r2),2),0)
plot(x,y,type="l", ylab="Density of t(2)",xlim=c(-5,5))
```

```
abline(h=0);polygon(x2,y2,col="yellow")
title(expression(
paste("Tail Probability for t(2) ",(alpha==0.05))))
text(c(0,4,3),c(.15,.03,.01),c(expression(P(T<t[0.05])==0.95),
expression(P(T>t[0.05])==0.05), expression(t[0.05]==2.92)))
```

图4.10是由下面R代码实现的:

```
x=seq(0,8,l=1000);y1=df(x,3,20);y2=df(x,50,20)
plot(x,y2,type="l",xlab="",ylab="")
lines(x,y1,lty=2)
text(c(0.1,1.8),c(0.75,1),c("F(3,20)","F(50,20)"))
```

图4.11是由下面R代码实现的:

```
plot(c(-.5,0,1,1.5),c(0,1,0,0),type="s",xlab="",ylab="")
abline(h=0,lty=2)
```

图4.12是由下面R代码实现的:

```
d= c(1,3,10,100);par(mfrow=c(2,2));for(i in d){
z=NULL; for(j in 1:600)z=c(z,mean(runif(i)));
hist(z,pr=T,main=substitute(n==that,list(that=i)),
    xlim=c(0,1),col=4)}
```

图4.13是由下面R代码实现的:

```
x=c(seq(-2,8,length=1000));y=dnorm(x,3,1.5)
r2=2;r1=-4;r3=4;x2=c(r1,r1,x[x<r2&x>r1],r2,r2)
y2=c(0,dnorm(c(r1,x[x<r2&x>r1],r2),3,1.5),0)
x3=c(r1,r1,x[x<r3&x>r1],r3,r3)
y3=c(0,dnorm(c(r1,x[x<r3&x>r1],r2),3,1.5),0)
plot(x,y,type="l",xlab="X Value",ylab="Density of N(3,1.5)")
title("P(2<X<4) for N(3,1.5) Variable X");abline(0,0)
polygon(x3,y3,col="grey");polygon(x2,y2,col="yellow")
text(c(1,2,2.5,3,4.5),c(.02,.25,.01,.15,.01),c("P(X<2)",
"P(X<4)","x=2","P(2<X<4)","x=4"))
```

4.7 习题

1. 如果一名嫌疑人的血液和犯罪现场留下的血液按照DNA分析只有十万分之一的可能不一样. 你如何判断和解释?

2. 如果有百分之五的人是左撇子, 而你和你兄弟都是左撇子. 那么你和你兄弟都是左撇子这样事件的概率是不是$0.05 \times 0.05 = 0.0025$? 为什么?

3. 一辆汽车的前灯在一年内失效的概率为0.2, 而该车的电池在一年中失效的概率为0.1. 那么这两项同时失效的概率是不是$0.2 \times 0.1 = 0.02$? 如果电池是另外车上的, 答案有所不同吗? 请用常识判断.

4. 在5个人中只有一张球票. 于是抽签决定谁去. 假定抽签是随机的. 机会应该均等. 但是你是最后一个抽, 如果前面没有人抽到, 你的机会不就是百分之百了吗? 而如果前面有人抽到, 你的机会不就是0了吗? 这样, 你还没有抽, 命运就已经决定. 这公平吗?

5. 每天你都会在上班路上遇到一些从未见过的人, 因此, 这显然是小概率事件. 但你又想, "天天都发生的事情会是小概率事件吗?"请和同学讨论这个问题.

6. 如果由你从0到9中随机抽取一个数算是一个试验, 重复这样的试验10次, 那么, 得到0147802493和得到9999999999的概率是否一样? 无论你怎么回答, 请给出这两个事件的概率.

7. 假定$p = 0.1$是每次Bernoulli试验中成功的概率. 使用计算机或者适当的分布表计算(如果你愿意, 也可以用公式)

   (a) 在15次试验中至少3次成功的概率;

   (b) 在10次试验中最多1次成功的概率;

   (c) 在12次试验中, 成功次数至少3次而最多5次的概率.

8. 假定X为$N(2,2)$分布, 使用计算机或者适当的分布表计算(如果你愿意, 也可以用公式)

   (a) X大于8的概率;

   (b) X小于0的概率;

   (c) X在7和8之间的概率;

   (d) X在1和2之间的概率;

   (e) X在-4和8之间的概率.

第五章　简单统计推断：总体参数的估计

人们每时每刻都在做估计. 出门根据天色云量等估计今天的天气, 根据婴儿的哭声和面色判断其冷热和是否饥饿, 根据望闻问切来估计病人的病情, 根据外表估计一个人的身高体重, 根据营业数据等估计一个公司的业绩等等. 估计就是根据你拥有的信息来对现实世界进行某种判断. 统计中的估计也不例外, 它是根据数据做出的.

举例说, 人们想知道到底有多大比例的北京人同意北京大力发展轨道交通, 由于不大可能询问所有的一千多万北京市民, 人们只好进行抽样调查以得到样本, 并用样本中同意发展轨道交通的比例来估计真实的比例. 从不同的样本得到的结论也不会完全一样. 虽然真实的比例在这种抽样过程中永远也不知道, 但有可能知道估计出来的比例和真实的比例大致差多少. 从数据得到关于总体参数的一些结论的过程就叫做**统计推断(statistical inference)**. 这个调查例子是估计总体参数(某种意见的比例)的一个过程. **估计(estimation)**是统计推断的重要内容之一. 统计推断的另一个主要内容是下一章要引进的**假设检验(hypothesis testing)**.

5.1　用估计量估计总体参数

总体代表人们所关心的那部分现实世界. 而在利用样本中的信息来对总体参数进行推断之前, 人们往往对代表总体的变量假定了分布族. 比如假定某特定人群的身高属于正态分布族, 或者在抽样调查时对某个观点认同与否假定了二项分布族等等. 这些模型假定基本上是根据经验而得, 所以仅仅是对现实世界的一个近似. 在假定了总体分布族之后, 进一步对总体的认识就是要在这个分布族中选择一个与人们所关心的问题有关的具体分布. 由于分布族成员是由参数确定的, 如果能够估计出参数, 对总体的具体分布就知道得差不多了.

哪些是分布的参数呢? 一些常见的参数包括总体均值(μ), 总体标准差(σ)和(Bernoulli试验中)成功概率p等(总体中含有某种特征的个体之比例). 正态分布族中的成员被(总体)均值和标准差完全确定, Bernoulli分布族的成员被概率(或比例)p完全决定. 因此如果能够对这些参数进行估计, 总体分布也就估计出来了.

估计当然要根据从总体所抽取的样本来确定. 前面提到过, 样本的(不包含未知总体参数的)函数称为统计量, 而用于估计的统计量称为**估计量(estimator)**. 由于一个统计量对于不同的样本取值不同, 所以, 估计量也是随机变量, 并有其分布. 当然, 如果样本已经得到, 把数据代入之后, 估计量就有了一个数值, 也就不是随机的了, 这个数字称为该估计量的一个**实现(realization)**或取值, 也称为一个**估计值(estimate)**.

这里介绍两种估计, 一种是**点估计(point estimation)**, 也就是用估计量的实现值来近似相应的总体参数. 另一种是**区间估计(interval estimation)**, 它是

包括估计量在内(有时是以估计量为中心)的一个区间, 该区间被认为很可能包含总体参数. 点估计给出一个数字, 用起来很方便, 而区间估计给出一个区间, 说起来留有余地, 不像点估计那么绝对.

思考一下:

1. 人们往往假定某感兴趣的总体有某种分布, 然后通过从这个总体抽出来的样本来得到这个总体分布参数的性质, 这就是统计推断中的估计. 然而, 并不是这些关于总体的假定都有道理, 因此, 统计中还有一些判断这些总体的近似分布的方法.

2. 现实世界大多数总体的实际分布是不可能知道的, 更不可能仅仅限于人类所掌握的少数可以用数学语言表达出来的那些. 人类所能够希望的是某些已知的概率分布可以对某些实际总体有较好的近似而已.

3. 人们根本不可能根据数据来证明该数据来自于哪个总体, 但可以收集证据来试图否定数据属于某个总体, 如果否定不了, 最多只能够说, 没有证据否定该数据来源于该总体. 第六章的假设检验就反映了这种科学的否定式思维方法.

5.2 点估计

用什么样的估计量来估计参数呢? 实际上没有硬性限制. 任何统计量, 只要人们觉得合适就可以当成估计量. 当然, 统计学家想出了许多标准来衡量一个估计量的好坏. 每个标准一般都仅反映估计量的某个方面. 这样就出现了按照这些标准定义的各种名目的估计量. 另一些估计量则是由它们的计算方式来命名的. 最常用的估计量就是大家熟悉的样本均值($\bar{X}$)、样本标准差(s)和(Bernoulli试验的)成功比例(x/n), 人们用它们来分别估计总体均值(μ)、总体标准差(σ)和成功概率(或总体中的比例)p. 这些在前面都已经介绍过, 大家也知道如何通过计算机(或公式)来计算它们.

那么, 什么是好估计量的标准呢? 一种统计量称为**无偏估计量(unbiased estimator)**. 所谓的**无偏性(unbiasedness)**就是: 虽然每个样本产生的估计量的取值不一定等于参数, 但当抽取大量样本时, 那些样本产生的估计量的均值会接近真正要估计的假定分布的参数. 严格说来, 如果估计量的数学期望等于欲估计的总体参数, 则该估计量称为该参数的无偏估计量. 注意, 由于一般仅仅抽取一个样本, 并且用该样本的这个估计量的实现来估计对应的参数, 人们并不知道这个估计值和要估计的参数差多少. 因此, 无偏性仅仅是非常多次重复抽样时的一个渐近概念. 随机样本产生的样本均值、样本标准差和Bernoulli试验的成功比例分别都是相应的总体均值、总体标准差和总体比例的无偏估计. 在无偏估计量的类中, 人们还希望寻找方差最小的估计量, 称为最小方差无偏估计量. 这是因为方差小说明反复抽样产生的许多估计值差别不大, 因此更加精确. 评价一个统计量好坏的标准很多, 而且许多都涉及一些大样本的极限性质. 我们不想在这里涉及太

多此方面的细节.

作为最小方差无偏估计的描述性例子, 假定$X_1, ..., X_n$ $(n > 2)$为来自一个总体的独立随机样本, 这些观测值互相独立, 那么, 对于总体均值μ的无偏估计就有很多. 比如下面的统计量都是无偏估计[1], 它们的期望都是μ.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad (X_1 + X_2)/2, \quad \frac{1}{3}X_1 + \frac{2}{3}X_2, \quad X_1,$$

但是, 他们的标准差不同, 第一个是$\sigma/\sqrt{n}$, 第二个是$\sigma/\sqrt{2}$, 第三个是$\sqrt{5/9}\sigma$, 而最后一个是σ. 显然$\bar{X}$的标准差最小, 当然方差也最小.

思考一下:

1. 日常生活中有许多估计, 有些类似于点估计. 比如, 人们会说某些商店贵, 某些商店便宜, 这肯定不是指一两个商品. 请讨论人们可能如何思考这个问题.
2. 举例说明你可以想象的点估计的应用.
3. 估计的精确性(如无偏性和最小方差性等等)仅仅是对人们所猜想的模型(或总体)而言, 但模型本身和现实世界的差距就不得而知了. 因此, 任何推断的精确性都无法弥补一个拙劣猜想的模型所造成的与实际问题之间的巨大偏差.

5.3 区间估计

当描述一个人的体重时, 你一般可能不会说这个人是76.45公斤, 而说这个人是七八十公斤, 或者是在70公斤到80公斤之间. 你提供的这个范围就是某种区间估计. 在抽样调查例子中也常用点估计加区间估计的说法. 比如, 为了估计某候选人在选民中的支持率(即总体比例p), 调查机构的民意测验可能会说, 该候选人的"支持率为75%, 误差是±3%, 置信度为95%". 这种说法意味着下面三点:

1. 样本中的支持率为75%, 这是用样本比例作为对总体比例的点估计.
2. 估计范围为$75\% \pm 3\%$(±3%的误差), 即区间$(72\%, 78\%)$.
3. 如果用类似的方式, 重复抽取大量(样本量相同的)样本时, 产生的大量类似区间中有些会覆盖真正的p, 而有些不会, 但这些区间中大约有95%会覆盖真正的总体比例.

这样得到的区间被称为总体比例p的**置信度(confidence level)**为95%的**置信区间(confidence interval)**. 这里的置信度又称**置信水平**或**置信系数**.

显然置信度的概念又是大量重复抽样时的一个渐近概念. 因此, "目前得到的区间(比如上面的$75\% \pm 3\%$)以概率0.95覆盖真正的比例p"是个错误的说法. 这里

[1]有兴趣的读者可参看4.6节的总体均值和方差的性质.

的区间(72%, 78%)是固定的, 而总体比例p也是固定的值, 只不过未知而已. 因此只有两种可能: 或者该区间包含总体比例, 或者不包含, 这当中没有任何概率可言. 至于区间(72%, 78%)是否覆盖真实比例, 除非一个不漏地调查所有选民, 否则永远也无法知道. 事实上, 本书涉及的置信区间(或其上下界)都是由统计量来确定的, 依样本而变, 是随机变量. 因此, 构造置信度$100(1-\alpha)\%$置信区间的**随机区间**则以$1-\alpha$的概率覆盖待估计的参数, 但该区间相应于一个样本的实现值就是固定的了, 无法知道其是否真正覆盖需要估计的参数.

在用语上, 人们喜欢用啰嗦的符号$100(1-\alpha)\%$来表示置信度的记号. 因此95%置信区间相当于$\alpha = 0.05$的情况. 当然, 这可能只是西方人把百分号%(percentage)作为名词的习惯而已. 对于中国读者, 95%和0.95或者$1-0.05$都是很自然的同义语. 现在中国(包括官方)也已经学会用"百分点"(%)这个名词(量词)了. 在英文中, 置信区间的上界(或上限, 即区间的左边界)称为upper bound, 下界(或下限)称为lower bound.

5.3.1 一个正态总体均值μ的区间估计

刚才所涉及例子是关于总体比例的置信区间, 本章后面还要给出计算公式. 除了比例之外, 还可以对其他参数, 例如总体均值构造置信区间. 下面看一个数值例子.

例5.1 (数据: noodle.txt) 某厂家生产的挂面包装上写明"净含量450克". 在用天平称量了商场中的48包挂面之后, 得到样本量为48的关于挂面重量(单位: 克)的一个样本:

```
449.5 461.1 457.5 444.7 456.1 454.7 441.5 446.0 454.9 446.2 457.3 446.1
456.7 451.4 452.5 452.4 442.0 452.1 452.8 442.9 449.8 452.4 458.5 442.7
447.9 450.5 448.3 451.4 449.7 446.7 441.7 455.6 442.9 451.3 452.9 457.2
448.5 444.5 443.1 442.3 439.6 446.5 447.2 445.8 449.4 441.6 444.7 441.4
```

这里假定, 挂面重量所代表的总体分布服从正态分布. 利用计算机, 可以很容易地得到挂面重量的样本均值及总体均值的置信区间等.

用下面R语句可以很容易得到关于该数据的各种常用统计量:

```
weight=scan("noodle.txt") #读入数据
summary(weight)           #输出均值,中位数,极大极小值,上下四分位点
t.test(weight,con=.95)$con#输出95%置信区间
```

输出的样本均值等于449.01, 而总体均值的95%置信区间为(447.41, 450.61). 这个置信区间是根据公式

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

计算出来的, 这里n是样本量, $\bar{x}$是样本均值的实现值, $t_{\alpha/2}$为自由度为$n-1$的t分布的上$\alpha/2$分位点, s是样本标准差. 当然, 根本不用麻烦地代公式手工计算(包括查表求$t_{\alpha/2}$), 用一句计算机代码很容易得到结果.

图5.1展示了一个模拟结果, 描述了从一个正态总体抽样得到的各种样本, 在不同置信度, 不同样本量的情况下关于总体均值置信区间长短和覆盖的情况. 这里虚线为"真实"的均值. 可以看出对于同样的样本量(这里分别是$n = 50$或$n = 20$), 置信度的增加导致区间变长. 对于同样的置信度(这里是0.95或0.60), 样本量的增加导致区间变短. 而无论样本量多少, 显然置信度大的, 覆盖真实总体均值的区间比例要大些.

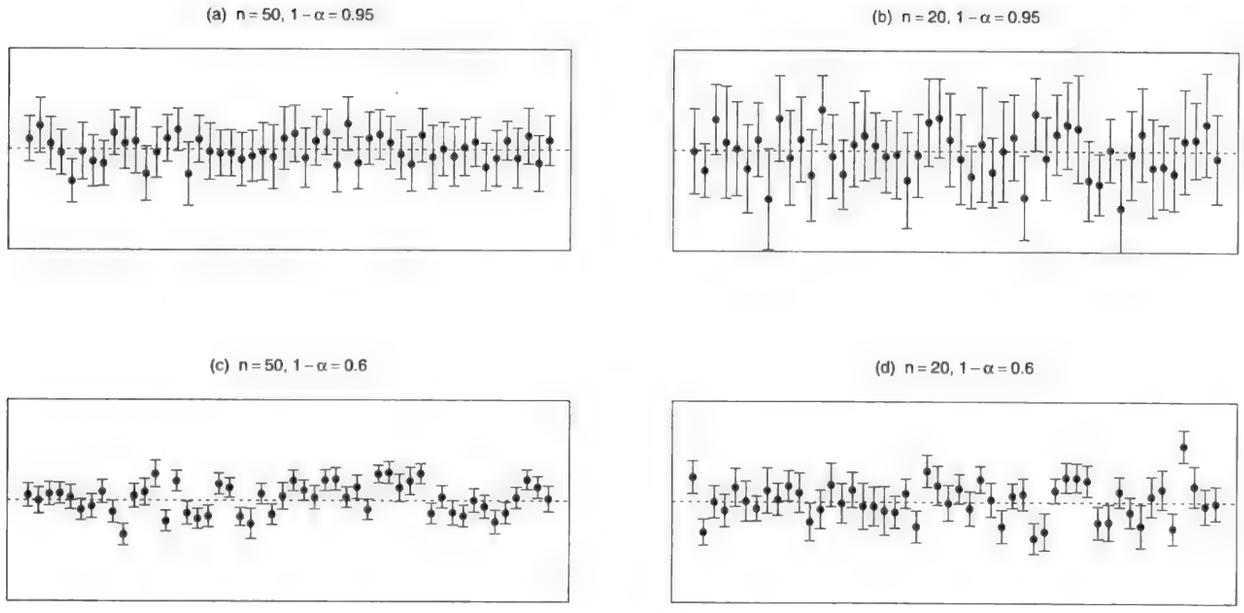

图 5.1 不同样本量和不同置信度的置信区间的长短和覆盖状况.

5.3.2 两个正态总体均值之差$\mu_1 - \mu_2$的区间估计

人们不仅可以构造一个总体均值(或比例)的置信区间, 还可以构造两个总体均值(或比例)之差的置信区间. 比如, 希望知道两个地区学生成绩的差异, 可以建造两个地区成绩均值之差$\mu_1 - \mu_2$的置信区间. 如果想要比较一个候选人在不同阶段支持率的差异, 那就可以构造两个比例之差$p_1 - p_2$的置信区间等. 下面再看一个例子.

例5.2 (数据: expend.txt) 这是我国两个地区的一些城市2003年的城镇家庭人均消费性支出(单位: 元)数据. 这里, 假定这种支出服从正态分布. 在数据中(无论哪种形式)收入是一列, 变量名为expend, 而区域为另一列, 变量名为area(用1、2表示两个地区).

(a) 希望要分别得到这两个总体均值和标准差的点估计(即样本均值和样本标准差)和各自总体均值的95%置信区间. 利用R语句:

```
w=read.table("expend.txt",header=T)#读入数据
```

```
x=w[w[,2]==1,1];y=w[w[,2]==2,1] #分开两个区域
mean(x);sd(x);mean(y);sd(y)#得到各自的均值和标准差:
```

作为两个总体均值估计量的样本均值分别为4562.53和5413.72, 而样本标准差分别为599.831和785.121.

(b) 求两个均值差$\mu_1 - \mu_2$的点估计和95%置信区间. 可用下面语句得到可得到均值差的置信区间:

```
mean(x)-mean(y)#均值差的估计
t.test(x,y)$conf
```

两个总体均值差$\mu_1 - \mu_2$的估计为-851.1928, $\mu_1 - \mu_2$的95%置信区间为$(-1330.8755, -371.5101)$. 注意, 这个结果假定了两个区域数据的方差不等. 如果假定两个区域数据方差相等, 则用(增加选项"var=T")代码

```
t.test(x,y,var=T)$conf
```

得到两个总体均值差$\mu_1 - \mu_2$的95%置信区间为$(-1333.8228, -368.5629)$. 如何判断方差是否相等呢? 这需要用下一章将讲到的检验. 这可以用R语句var.test(x,y)$p.value实行, 得到一个$p$值(下一章介绍), 这里是0.2880653. 如果p值很小(一般认为小于0.05, 但不是绝对的), 称为显著, 认为方差不相等. 这里的p值并不小, 所以, 没有证据说明方差不相等, 但**绝对不等于证明了方差相等**.

思考一下:

1. 上面的输出中, 必须先检验一下方差, 再在两行中挑选一行看结果, 有些麻烦. 实际上, 如果相信数据, 直接用根据数据中提供的信息而建立的第二行结果, 也未尝不可. 结果差不了多少. 但是由于这里第一行假定方差相等的计算公式是先有的, 后来才有的更一般的第二行计算公式. 所以, 可能是为了尊重历史. 实际上, 任何两个总体的方差都不可能完全相同, 如果真的类似, 用第二行结果也差不了多少. R软件的默认值就是方差不相等的情况.
2. 本书中的数值结果包含的小数点后的位数往往过多, 这是直接从计算机上复制下来的结果, 为了比较, 本书大多不刻意地减少小数点后的位数.

5.3.3 总体比例(Bernoulli试验成功概率)p的区间估计

例5.3 在一个大都市中对1341人的随机调查结果显示, 有934个人支持绿色出行和发展公共交通的政策. 假定该样本为简单随机样本, 希望找出总体中支持绿色出行和发展公共交通的人的比例的点估计及其置信度为95%的置信区间.

首先, 由于总体很大, 该调查可以看成是Bernoulli试验, 而支持绿色出行和发展公共交通的总体比例p的点估计可以很容易算出: $\hat{p} = 934/1341 = 0.6964952$. 而使用R语句

```
binom.test(934,1341,con=.95)$con
```

可以得到p的95%置信区间为$(0.6711031, 0.7210230)$. 下面语句可以算出包括精确区间在内的p的三种区间及其点估计(要事先下载程序包Hmisc[1]). 输出中的渐近(asymptotic)区间是用Bernoulli试验的大样本正态近似的置信区间的公式计算的, Wilson区间是正态近似区间的改进.

```
library(Hmisc); binconf(934, 1341, alpha=.05,method="all")
```

输出为

```
           PointEst     Lower     Upper
Exact     0.6964952 0.6711031 0.7210230
Wilson    0.6964952 0.6713547 0.7205131
Asymptotic 0.6964952 0.6718872 0.7211031
```

之所以有各种置信区间的算法, 是因为在前计算机时代, 算精确区间很不容易, 人们就用各种方法来寻找计算量少的近似区间.

注意, 这里的方法仅限于大总体的情况. 如果用正态近似, 则只适用于大总体及大样本. 在小总体和小样本时要用超几何分布的模型, 而在大总体和小样本时不能用大样本正态近似, 必须直接用精确方法. 读者可参看后面公式.

5.3.4 总体比例(Bernoulli试验成功概率)之差$p_1 - p_2$的区间估计

例5.4 在两个地区对于某商品认可与否的调查结果显示, 第一个地区被调查的950人中有423人认可, 而在第二个地区的被调查的1102人中只有215人认可. 求这两个总体比例之差$p_1 - p_2$的95%置信区间.

用一句R代码

```
prop.test(c(423,215),c(950,1102),con=.95)$con
```

可以得到两个总体比例之差$p_1 - p_2$的95%置信区间为(0.2098615, 0.2904652).

5.4 关于置信区间的注意点

前面已经提到, 不要认为由一个样本数据得到总体参数的一个95%置信区间, 就以为该区间以0.95的概率覆盖总体参数. 置信度95%仅仅描述用来构造该区间上下界的统计量(是随机的)覆盖总体参数的概率, 也就是说, 无穷次重复抽样所得到的所有区间中大约有95%包含参数. 但是把一个样本数据带入统计量的公式所

[1]Frank E Harrell Jr and with contributions from many other users (2012). Hmisc: Harrell Miscellaneous. R package version 3.9-3. http://CRAN.R-project.org/package=Hmisc.

得到的一个区间, 只是这些区间中的一个. 这个非随机的区间是否包含那个非随机的总体参数, 一般不可能知道. 非随机的数目之间没有概率可言.

置信区间的论述是由区间和置信度两部分组成. 有些新闻媒体报道的一些调查结果只给出百分比和误差(即置信区间), 并不说明置信度, 也不给出被调查的人数, 这是不负责的表现. 因为降低置信度可以使置信区间变窄(显得"精确"), 有误导读者之嫌. 在公布调查结果时给出被调查人数是负责任的表现. 这样内行则可以由此推算出置信度(由后面给出的公式), 反之亦然.

一个描述性例子: 一个有10000个人回答的调查显示, 同意某种观点的人的比例为70%(有7000人同意), 可以算出总体中同意该观点的比例的95%置信区间为(0.691, 0.709)(用代码`binom.test(7000,10000,con=.01)$con`), 另一个调查声称有70%的比例反对该种观点, 还说总体中反对该观点的置信区间也是(0.69, 0.71). 到底相信谁呢? 实际上, 第二个调查隐瞒了置信度(等价于隐瞒了样本量). 如果第二个调查仅仅调查了50个人, 有35个人反对该观点. 可以算出, 第二个调查的置信区间的置信度仅有1%(用代码`binom.test(35,50,con=.01)$con`).

常识表明, 来自现实世界的数据量越大, 对现实世界的了解就越充分. 样本量对置信区间有很大的影响. 理想的情况是获得很小的置信区间和很大的置信度. 但鱼与熊掌不可兼得, 只好固定一个, 力求另一个更好. 如果固定置信度在某个值, 比如95%, 那么样本量越大, 置信区间就越窄. 如果固定置信区间的长度, 那么样本量越大, 置信度就越大. 人们可以从需要的置信区间的长度和置信度求出需要多大的样本量. 当然, 要指明的是, 在固定置信度时, 置信区间长度的减少并不是和样本量n成反比, 而是和$\sqrt{n}$成反比, 也就是说当样本量增加一倍(即$2n$)时, 置信区间的长度为原先的$1/\sqrt{2}$.

这里所涉及的一些区间估计的公式在后面会介绍, 同时还会总结如何使用R软件从数据获得想要的区间估计.

思考一下:

1. 对于正态总体, 在样本没有得到时, 端点由随机的统计量组成的诸如$\bar{x} \pm t_{\alpha/2}s/\sqrt{n}$的$100(1-\alpha)\%$ 置信区间的确以$1-\alpha$的概率覆盖真实总体均值, 但一旦得到样本数据, 并以此计算出一个具体数值区间, 比如例5.1中的95%置信区间(447.41, 450.61). 这时, 就不能说区间(447.41, 450.61)以0.95的概率覆盖均值. 为什么这样, 请讨论.

2. 置信区间的概念需要对总体有所要求或假定, 请讨论.

5.5 小结

5.5.1 本章的概括和公式

本章的内容很简单, 就是作为统计推断重要经典内容的估计, 包括点估计和区间估计. 下面介绍进行计算所依据的有关公式.

1. 一个正态总体均值μ的置信区间

假定独立观测值$x_1, ..., x_n$形成服从正态分布的样本量为n的一个样本. 那么总体均值μ的$100(1-\alpha)\%$ 置信区间为

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \quad \text{或者} \quad \left(x - t_{\alpha/2}\frac{s}{\sqrt{n}},\ x + t_{\alpha/2}\frac{s}{\sqrt{n}}\right),$$

这里$\bar{x}$为样本均值, s为样本方差, $t_{\alpha/2}$为自由度为$n-1$的t分布相应于尾概率$\alpha/2$的t变量的值, 即对于自由度为$n-1$的t分布变量T, 有$P(T > t_{\alpha/2}) = \alpha/2$. 尾概率的概念和$t_{\alpha/2}$的求法在第四章已经介绍了. 在某些情况下, 假定了总体标准差σ是已知的, 这时, 总体均值μ的$100(1-\alpha)\%$置信区间为

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{或者} \quad \left(x - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ x + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right),$$

这里$\bar{x}$为样本均值, σ为已知的总体方差, $z_{\alpha/2}$为标准正态分布相应于尾概率$\alpha/2$的z变量的值. 这种σ已知的情况虽然少见, 但反映了置信区间的历史, 所以很多教科书都论及. 按照这两个公式以及前面两章提到的利用软件计算相应值的方法, 可以很容易的得到置信区间. 实际上, 前面的例子说明, 统计软件可以直接从数据把置信区间以及各种估计量算出. 不用按照公式分别计算.

2. 正态总体均值之差$\mu_1 - \mu_2$的置信区间

这里分总体方差相等和不相等两种情况. 实际上, 根本无法根据数据证明两个总体方差相等. 用下章要介绍的假设检验可以拒绝方差相等. 但有人觉得可以用证据不足以拒绝方差相等的假设来说明两个总体方差相等. 这是完全错误的. 因为在小样本时, 基本上都无法拒绝方差相等的假设, 这只能说证据不足, 因而仍然把方差相等作为一个数学假定(而不是事实!)

假定两总体方差相等假定下的公式. 假定独立观测值$x_1, ..., x_n$和$y_1, ..., y_n$ 形成两个服从正态分布的样本, 样本量分别为n_1和n_2, 那么两个总体方差相等的正态总体均值之差$\mu_1 - \mu_2$的$100(1-\alpha)\%$置信区间为

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

或者

$$\left((\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

这里$\bar{x}_1$和$\bar{x}_2$分别为两个样本均值, $t_{\alpha/2}$为自由度为$n_1 + n_2 - 2$的t分布相应于上侧尾概率$\alpha/2$的t变量的值, 即对于自由度为$n_1 + n_2 - 2$的t分布变量T, 有$P(T > t_{\alpha/2}) = \alpha/2$, 式中的$s_p$定义为

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 1}},$$

而s_1和s_2分别为两个样本的标准差.

假定两总体方差不相等假定下的公式. 假定独立观测值$x_1, ..., x_n$和$y_1, ..., y_n$形成两个服从正态分布的样本, 样本量分别为n_1和n_2, 那么两个总体方差不相等的正态总体均值之差$\mu_1 - \mu_2$的$100(1-\alpha)\%$置信区间为

$$(\bar{x}_1 - \bar{x}_2) \pm t^*_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

或者

$$\left((\bar{x}_1 - \bar{x}_2) - t^*_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},\ (\bar{x}_1 - \bar{x}_2) + t^*_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right)$$

这里$\bar{x}_1$和$\bar{x}_2$分别为两个样本均值, $t^*_{\alpha/2}$为近似自由度为ℓ的t分布相应于上侧尾概率$\alpha/2$的t变量的值, 即对于自由度为ℓ的t分布变量T, 有$P(T > t_{\alpha/2}) = \alpha/2$, 自由度$\ell$定义为

$$\ell = \frac{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}{\dfrac{s_1^2}{n_1^2(n_1-1)} + \dfrac{s_2^2}{n_2^2(n_2-1)}},$$

而s_1和s_2分别为两个样本的标准差.

3. Bernoulli试验参数(成功概率或总体比例) p的置信区间

(1) 大总体和大样本情况

对于Bernoulli试验中参数(成功概率)p的估计的最常见的例子是抽样调查中持某种观点的比例. 假定现在总体很大. 共调查了n个人(大总体时, 可以近似看为n次Bernoulli试验), 其中持某种观点的为x("成功"数目x), 于是样本比例为$\hat{p} = x/n$. 那么比例p的$100(1-\alpha)\%$近似置信区间为(这里是大样本正态近似, 因此与正态分布发生了关系, 我们用R软件可以得到精确区间和近似区间)

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad \text{或者} \quad \left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

警告: 这个公式一定要在大样本时才能够用. 什么是大样本呢? 一个简单的近似判别方法(仅仅是必要条件)是当区间

$$\hat{p} \pm 3\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

完全包含在(0, 1)区间内部时, 可以认为样本足够大, 能够用正态近似.

(2) 大总体和小样本情况

在大总体, 但小样本时有没有精确的关于比例(或近似的Bernoulli试验的成功概率)的置信区间的求法呢? 当然有, 只不过许多教科书不介绍、一些傻瓜软

件不支持而已(但R软件支持). 其基本思想如下. 用第四章的记号. 以$p(k)$代表在n次Bernoulli试验中成功的次数的概率, p为每次试验成功的概率. 有

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, ..., n.$$

这里

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

为二项式系数. 也常用C_n^k来表示. 如果已经观测到n次试验有x次成功, 那么p的$100(1-\alpha)\%$置信区间(p_L, p_U)的上限p_U则应该为满足

$$\sum_{k=0}^{x} p(k) = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k} = \frac{\alpha}{2}$$

的p, 而置信区间的下限p_L则应该为满足

$$\sum_{k=x}^{n} p(k) = \sum_{k=x}^{n} \binom{n}{k} p^k (1-p)^{n-k} = \frac{\alpha}{2}$$

的p. 除了R软件直接计算之外, 这可以通过二项分布的表, 或者软件中关于二项分布的累积分布函数做几次尝试而得. 当然, 如果软件中有二项分布的累积分布函数的逆函数则更方便了. 只要编一个一行的小程序就可得到结果.

(3) **小总体情况**

在小总体的抽样调查中求比例的问题大都应该属于超几何分布, 这是因为在调查中的抽样均属于不放回抽样. 由于一切统计模型都是近似模型, 超几何分布也不例外. 它要求总体中每一个个体有同等机会被抽到, 而这不可能在实践中完全做到. 作为超几何分布, 就应该有直接计算其置信区间的精确方法. 按照该方法, 这个置信区间应该从求k(比如总体中的废品个数)的$100(1-\alpha)\%$的置信区间着手, 而该区间(k_1, k_2)上限k_2应该为满足

$$P(N, n, k, x) \leq \frac{\alpha}{2}$$

的最小的k, 而其下限k_1应该为满足

$$P(N, n, k, x-1) \geq 1 - \frac{\alpha}{2}$$

的最大的k. 这里$P(N, n, k, x) \equiv P(X \leq x)$为参数为$N, n, k$的超几何分布的累积分布函数,

$$P(N, n, k, x) = \sum_{i=0}^{x} p(N, n, k, i),$$

而

$$p(N,n,k,x)=\frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}};\ \max[0,n-(N-k)]\le x\le\min(k,n)$$

有了区间(k_1,k_2)之后, 除以N就可以得到比例k/N的置信区间了.

注意, 在概念上, 如果抽样时总体中所有个体的等可能性可以保证, 则按超几何分布计算的精确区间可被看成为**精确模型的精确解**. 而在总体量大时按二项分布计算的精确区间为**近似模型的精确解**. 如果总体量比较大, 同时样本量也比较大(至少核对前面的近似必要条件), 才可以用二项分布的大样本近似求置信区间, 这时的解为**近似模型的近似解**.

4. **两个Bernoulli试验参数(成功概率或比例) 之差p_1-p_2的置信区间**

假定两个Bernoulli试验次数分别为n_1和n_2, 而成功比例次数分别为$\hat{p}_1=x_1/n_1$和$\hat{p}_2=x_2/n_2$, 那么, 参数之差p_1-p_2 的$100(1-\alpha)\%$置信区间为

$$(\hat{p}_1-\hat{p}_2)\pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

同样这个公式也只有在两个样本都足够大时才可以应用.

5.5.2 R语句的说明

本章所提到的连续变量总体均值和标准差的估计量就是在第三章提到的样本均值和样本标准差. 而总体比例的估计就是样本比例(简单的整数相除). 因此下面不赘述. 稍微复杂的是区间估计. 但都可以直接用R软件得出:

1. **单样本的总体均值μ的区间估计**: 对于数据代码x, 用R语句

```
t.test(x,conf=0.95)$conf
```

得到(95%)置信区间.

2. **两样本的总体均值差$\mu_1-\mu_2$的区间估计**: 对数据代码x和y, 用语句

```
t.test(x,y, conf=0.95)$conf
```

就可以得到(95%)置信区间. 如果想用前计算机时代方法(假定方差相等), 则先要用R语句`var.test(x,y)$p.value`得到p值, 如果p值不是很小, 可用语句`t.test(x,y, var=T,conf=0.95)$conf`得到假定方差相等的简单公式计算的(95%)置信区间.

3. **单样本的总体比例p的区间估计**: 对于数据代码n和x, 用R语句

```
binom.test(x,n,con=.95)$con
```

得到(95%)置信区间. 或者用

```
library(Hmisc); binconf(x, n, alpha=.05,method="all")
```

得到各种精确和近似置信区间.

4. **两样本总体比例差$p_1 - p_2$的区间估计**: 对数据代码x1, x2和n1, n2, 用语句

```
prop.test(c(x1,x2),c(n1,n2),con=.95)$con
```

就可以得到(95%)置信区间.

5.6 习题

1. 说出点估计和区间估计的不同以及各自的优缺点.

2. 如果一条广告说, 某药品的有效率为80%, 其误差为正负3%, 那么这条广告给出了什么信息? 你相信这条广告吗? 这条广告的发布者隐瞒了什么信息?

3. 如果在置信度不变的情况下, 你要使目前所得到的置信区间的长度减少一半, 样本量应增加到目前样本量的多少倍? 如果保持置信区间长度不变, 样本量增加会使什么变化?

4. 利用任何你觉得可用的方法(比如利用公式、查表或任何计算机软件)重复例2.

5. 如果得到均值的一个95%置信区间为(3.5, 4.3), 是否可以说区间(3.5, 4.3)以95%的概率覆盖总体均值? 是不是也可以说总体均值以95%的概率落入区间(3.5, 4.3)之中? 为什么? 怎样才是合适的说法?

6. 有一个商店雇员问了10个顾客是否喜欢该商店的服务, 结果是有7个人说喜欢. 于是该雇员根据公式$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$ 得到喜欢该商店服务的顾客比例的95%置信区间为(0.42, 0.98). 这样做有什么不妥吗?

第六章　简单统计推断: 总体参数的假设检验

在航天火箭发射前, 没有任何人能够事先证明火箭发射是安全的, 人们最多只能够说, 用现有手段没有发现问题. 但是, 只要发现一个影响安全发射的问题, 那就是问题, 火箭就不能发射. 这说明, 企图肯定什么事物很难, 而否定却要相对容易得多. 物理学以及其他科学都是在否定中发展的, 这也是假设检验背后的哲学. 在所有学科中, 只有纯粹数学是在一定公理系统下依赖肯定式演绎思维发展的, 各种科学都是在一个接一个地根据观测或实验结果否定各种假说或者理论的基础上发展的.

假设检验是一种方法, 目的是为了判断一个关于总体特征的定量的断言(假设)的真实性. 人们通过从总体中抽出的随机样本来计算适当的统计量来检验一个假设. 如果得到的统计量的实现值在假设为真时应该是罕见的(小概率事件), 则有理由拒绝这个假设.

在假设检验中, 一般要设立一个原假设(上面的"火箭发射是安全的"就是一个例子), 而设立该假设的动机主要是企图利用人们掌握的反映现实世界的数据来找出假设与现实之间的矛盾[1], 从而否定这个假设, 并称该检验**显著(significant)**. 在多数统计实践中(除了理论探讨之外)的假设检验都是以否定原假设为目标. 如果否定不了, 那就说明证据不足, 无法否定原假设. 但这不能说明原假设正确. 就像用一两个仪器没有发现火箭有问题还远不能证明火箭是安全的那样.

本章主要讨论关于连续变量总体均值和基于二项分布的总体比例的假设检验.

6.1　假设检验的过程和逻辑

6.1.1　假设检验的过程和逻辑

例6.1(数据: sugar.txt) 一个顾客买了一包标有500g重的红糖, 觉得份量不足, 于是找到监督部门, 当然他们会觉得一包份量不够可能是随机的. 于是监督部门就去商店称了50包红糖, 得到样本均值(平均重量)是498.35g, 这的确比500g少, 但这是不是仅仅是由于随机误差造成的呢? 这是否能够说明厂家生产的这批红糖平均起来不够份量呢? 首先, 可以画出这些重量的直方图(图6.1). 这个直方图看上去像是正态分布的样本. 于是不妨假定这一批袋装红糖呈正态分布[2].

图6.1是用下面R语句画的:

```
weight=scan("sugar.txt")#读入数据
hist(weight,main="Histogram of Sugar Weight")
```

首先要提出一个原假设, 比如例6.1的红糖重量问题, 原假设可为均值等于500g($\mu = 500$). 这种原假设也称为**零假设(null hypothesis)**, 记为H_0. 与此

[1]这里所谓的矛盾, 就是按照原假设, 现实世界数据的出现仅仅属于小概率事件, 是不大可能出现的.

[2]这种假定并不是一定成立的. 后面将介绍关于正态性的假设检验. 就这个例子而言, 常用的正态性检验(比如Shapiro-Wilk正态性检验)没有足够证据来拒绝该数据的正态性.

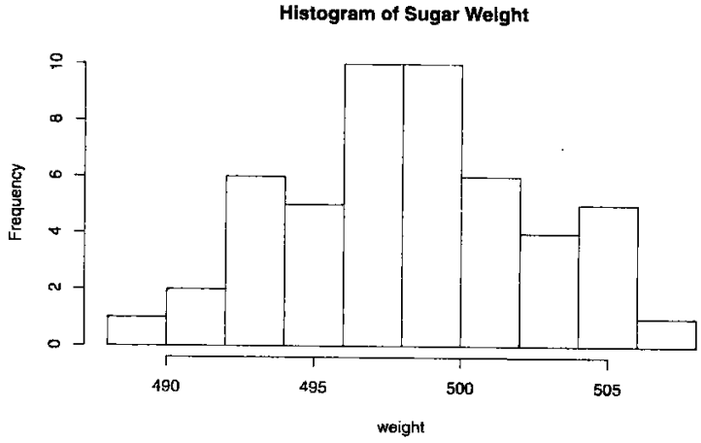

图 6.1 例6.1的50包红糖重量的直方图.

同时必须提出**备选假设**(或称为**备择假设, alternative hypothesis**), 比如总体均值小于500g($\mu < 500$). 备选假设记为H_1或H_a. 形式上, 上面的关于总体均值的H_0相对于H_1的检验记为

$$H_0: \ \mu = 500 \ \Leftrightarrow \ H_1: \ \mu < 500$$

这里符号"⇔"就是相应于英文versus, 类似于甲队对乙队比赛的"对"字. 备选假设的不等式应该按照实际数据所代表的方向来确定, 即它通常是被认为可能比零假设更加符合数据所代表的现实. 比如上面的H_1为$\mu < 500$, 这意味着, 至少样本均值应该小于500, 至于是否显著, 依检验结果而定. 检验结果显著意味着有理由拒绝零假设. 因此, 假设检验也被称为**显著性检验(significant test)**.

有了两个假设, 就要根据数据来对它们进行判断. 数据的代表是作为其函数的统计量, 它在检验中被称为**检验统计量(test statistic)**. 根据**零假设**(不是备选假设!)就可以得到该检验统计量的分布, 然后再看这个统计量的数据实现(realization)属不属于小概率事件出现了. 也就是说把数据代入检验统计量, 看其值是否落入零假设下的小概率范畴, 如果的确是小概率事件, 那么就有可能拒绝零假设, 或者说"该检验显著", 否则说没有足够证据拒绝零假设, 或者说"该检验不显著".

注意: 在本书所涉及的问题中, **零假设和备选假设在假设检验中并不对称**. 因检验统计量的分布是从零假设导出的, 因此, 如果发生矛盾, 就对零假设不利了. 不发生矛盾也不能说明零假设没有问题, 只能说证据不足以拒绝零假设.

在零假设下, 检验统计量取其实现值及(沿着备选假设的方向)更加极端值的概率称为***p*值(p-value)**. 为了说明上面定义的含义, 就本章将涉及的单边和双边检验问题而言, 假定某检验统计量T的样本实现值为t. 如果T越大就越有利于备选假设, 则p值等于零假设下统计量T取其实现值及更极端值的概率$P_{H_0}(T \geq t)$; 类似地, 如果T越小就越有利于备选假设, 则p值等于$P_{H_0}(T \leq t)$; 而如果绝对值$|T|$越大就越有利于备选假设, 则p值等于$P_{H_0}(|T| \geq |t|)$. 可以看出, p值和检验

统计量的实现值以及备选假设的方向有关. 如果得到很小的p值, 就意味着在零假设下小概率事件发生了. 如果小概率事件发生, 是相信零假设, 还是相信数据呢? 当然多半是相信数据. 于是就拒绝零假设. 但在零假设正确时, 小概率事件也还是可能发生, 仅仅是发生的概率很小罢了. 拒绝正确零假设的错误常被称为**第一类错误(type I error)**. 犯第一类错误的概率可以认为等于p值, 或者不大于(马上要介绍的)事先设定的显著性水平α.

那什么是第二类错误呢? 那是备选假设正确时没能拒绝零假设的错误, 称为**第二类错误(type II error)**. 在本书的假设检验问题中, 由于备选假设不是一个点, 所以无法算出犯第二类错误的概率.

另一个概念就是检验的**势(power)**, 对于统计学家来说, 检验的势就是当备选假设正确时, 该检验拒绝零假设的概率. 强势检验也称为高效率检验. 检验的势越强越好.

零假设和备选假设哪一个正确, 是确定性的, 没有概率可言. 而可能犯错误的是人. 涉及假设检验的犯错误的概率就是犯第一类错误的概率和犯第二类错误的概率. 负责任的态度是无论做出什么决策, 都应该给出该决策可能犯错误的概率.

到底p值要多小时才能够拒绝零假设呢? 也就是说, 需要有什么是小概率的标准. 这要看具体应用的需要. 但在一般的统计书和软件中, 使用最多的标准是在零假设下(或零假设正确时)根据样本所得的数据来拒绝零假设的概率应小于0.05, 当然也可能是0.01, 0.005, 0.001等等. 这种事先规定的概率称为**显著性水平(significant level)**, 用字母α来表示. α并不一定越小越好, 因为这很可能导致不容易拒绝零假设, 使得犯第二类错误的概率增大. 当p值小于或等于α时, 就拒绝零假设. 所以, α是所允许的犯第一类错误概率的最大值. 当p值小于或等于α时, 就说这个检验是显著的. 无论统计学家用多大的α作为显著性水平都不能脱离实际问题的背景. 统计显著不一定等价于实际显著. 反过来也一样.

实际上, 多数计算机软件仅仅给出p值, 而不给出一个确定的α. 这有很多方便之处. 比如$\alpha = 0.05$, 而假定所得到的p值等于0.001. 这时如果采用p值作为新的显著性水平, 即新的$\alpha = 0.001$, 于是就可以说, 在显著性水平为0.001时, 拒绝零假设. 这样, 拒绝零假设时犯错误的概率实际只是千分之一而不是旧的α所表明的百分之五. 在这个意义上, p值又称为**观测的显著性水平(observed significant level)**. 在统计软件输出p值的位置, 有的用"p-value", 有的用significant的缩写"Sig"就是这个道理. 根据数据产生的p值来减少α 的值以展示结果的精确性总是没有害处的. 这好比一个身高180厘米的男生, 可能愿意被认为高于或等于180厘米, 而不愿意说他高于或等于155厘米, 虽然这第二种说法数学上没有丝毫错误.

在前计算机时代, 在假设检验中从若干显著性水平中选择是因为无法计算p值, 而教科书及手册只能提供相应于有限α值的表格. 这时, α取值为0.05, 0.01, 0.005, 0.001等简洁形式的值也是很自然的. 不能想象一个表格是用$\alpha = 0.009753$之类的显著性水平制成. 但今天的p值则可能是任何非负值, 因此, 一些现在的教科书还是引进满足人们某种心理的α值来和不那么整洁的p值比较.

在一些中国出版的统计教科书中会有不能拒绝零假设就"接受零假设"的说法[1]. 这种说法是不严格的. 首先, 如果你说"接受零假设", 那么就应该负责任地提供接受零假设时可能犯第二类错误的概率. 这就要算出在备选假设正确的情况下错误地接受零假设的概率. 但是, 这只有在备选假设仅仅是一个与零假设不同的确定值(而不是范围)时才有可能. 多数基本统计教科书的备选假设是一个范围, 例如在前面例子提到的检验问题$H_0: \mu = 500 \Leftrightarrow H_1: \mu < 500$ 的情况. 这时根本无法确定犯第二类错误的概率. 在许多诸如应用回归分析等领域的教科书中, 也往往把一系列不能拒绝零假设的检验当成接受这些假设的通行证. 比如不能拒绝某样本的正态性就变成了证明了该样本是正态的等等. 其实, 不能拒绝这些零假设, 仅仅说明根据所使用的检验方法(或检验统计量) 和当前的数据没有足够证据拒绝这些假设而已. 对于同一个假设检验问题, 往往都有多个检验统计量, 而且人们还在构造更优良的检验统计量. 人们不可能把所有目前存在的和将来可能存在的检验都实施. 因此, 在不能拒绝零假设时, 只能够说, 按照目前的证据和检验方法, 不足以拒绝零假设而已, 而零假设仍然是一个数学假定. 后面将会用例子说明"接受零假设"的说法是不妥当的. **统计工作者必须给用户一个没有偏见的信息, 而不是代替用户做没有指明风险的决策.**

思考一下:

1. 如果零假设是"某人从来不骂人", 要证明这一假设容易吗? 但只要发现其骂过一次, 这个假设就可以否定了. 请讨论科学研究中的类似情况.
2. 假定你抓住一个刚把赃物扔掉的小偷, 但没有赃物不能证据说明他(她)不是小偷, 只能够说明证据不足. 这和在无法在证据不足时不能说"接受零假设"时有同样的逻辑.

6.1.2 假设检验在前计算机时代发展的一些概念和步骤

1. 假设检验的逻辑步骤

在前计算机时代的课本都会列出下面的假设检验的步骤:

(1) **写出零假设和备选假设.** 比如对于总体均值的检验, 零假设为企图拒绝的量, 而备选假设需要看样本均值和零假设均值的相对大小来定.

(2) **确定检验统计量.** 本章都是常用的一些统计量(在计算机时代则选择检验方法, 计算机会自动按相应公式计算.).

(3) **确定显著性水平α.** 这个是你自己根据实际问题的需要来确定. 在前计算机时代, 只能在几个有限值中挑选. 在计算机时代则在下面p值确定之后决定.

[1]在国外最近三十年出版的统计教科书中未发现有(在没有给出犯第二类错误概率的情况下) "接受零假设"的说法. 而在中国, 过去四五十年出版的教材, 特别是一些"权威"教材, 不乏"接受零假设"的说法, 这可能是一种有中国特色的"习惯"或"传统"吧.

(4) **根据数据计算检验统计量的实现值.** 这一步过去要代公式计算(现在计算机会代劳).

(5) **得到检验是否显著的结论:**

(a) 在前计算机时代, 用实现值来比较根据α查表得到的"临界值"(下面有说明), 如果进入了临界值的"否定域"则认为检验显著, 拒绝零假设.

(b) 在计算机发展的今天, 计算机会根据实现值计算p值. 如果p值小于或等于α则认为检验显著, 拒绝零假设.

注意: 当上面的第(1)款确定之后, 其余皆由计算机自动完成. 这里所列出的几条, 是前计算机时代手工计算的思维和运作步骤.

2. **关于"临界值"的注**

作为概率的显著性水平的α实际上相应于一个检验统计量(比如T)取值范围的一个**临界值(critical value)**(这里暂时用t_α表示), 它定义为, 统计量取**该值或更极端的值**的概率等于α(比如, $P_{H_0}(T \geq t_\alpha) = \alpha$, $P_{H_0}(T \leq t_\alpha) = \alpha$或$P_{H_0}(|T| \geq |t_\alpha|) = \alpha$, 依备选假设的方向而定). 也就是说, **"统计量的实现值比临界值更极端"等价于"p值小于α"**. 使用临界值的概念进行的检验不计算p值. 只比较统计量的取值($T = t$)和临界值t_α的大小. 统计量的实现值比临界值更极端的取值范围也称为**"拒绝域"**.

以例6.1为例, 如果设定显著性水平为$\alpha = 0.005$, 那么, 对于自由度为49的t分布相应的临界值为$t_\alpha = -2.679952$(这不是查表得到的, 而是用R语句`qt(.005,49)`算出的), 因此, p值小于0.005等价于检验统计量的值(这里是-2.696)比t_α还要极端, 即小于t_α, 这时拒绝域为$(-\infty, -2.679952)$.

使用临界值而不是p值来判断拒绝与否是前计算机时代的产物. 当时计算p值不易, 只采用临界值的概念. 但从给定的α求临界值同样也不容易, 好在习惯上仅仅在教科书中列出相应于特定分布的几个有限的α(比如$\alpha = 0.05, \alpha = 0.025, \alpha = 0.01, \alpha = 0.005, \alpha = 0.001$等等)的临界值, 或者根据分布表反过来查临界值(很不方便也很粗糙). 现在计算机软件大都不给出α和临界值, 但都给出p值和统计量的实现值, 让用户自己决定显著性水平是多少.

显著性水平和临界值的概念都出现于前计算机时代, 但一些教科书还延用至今, 主要企图说明假设检验的逻辑过程. 那时的检验方向、显著性水平(临界值)的确定都是在抽样之前决定的, 但现在(至少在本书涉及到的检验中)则以数据为准, 一般有了数据才确定检验方向, 并根据数据算出p值来做出最后关于检验的决策.

6.2 对于正态总体均值的检验

6.2.1 根据一个样本对其总体均值大小进行检验

假定一个样本来自于均值为μ的正态总体, 人们想检验这个均值是否等于一个确定的数目, 比如说μ_0. 这就可以利用下面的t检验来实现. 继续看例6.1.

例6.1(数据: sugar.txt, 继续) 监督部门称了50包标有500g重的红糖, 均值是498.35g, 少于所标的500g. 对于厂家生产的这批红糖平均起来是否够份量, 需要统计检验. 由于厂家声称每袋500g, 因此零假设为总体均值等于500g(被怀疑对象总是放在零假设), 而且由于样本均值少于500g(这是怀疑的根据), 把备选假设定为总体均值少于500g(这种备选假设为单向不等式的检验称为单尾检验, 而备选假设为不等号"$\neq$"的称为双尾检验, 下面会解释). 即

$$H_0: \ \mu = 500 \ \Leftrightarrow \ H_1: \ \mu < 500,$$

而检验统计量就是第四章引进的作为对均值的某种标准化的

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

符号中的μ_0通常表示为零假设中的均值(这里是500). 在零假设之下(体现在公式中的μ_0), 根据例6.1对总体的正态性假定, 它有自由度为$n-1=50-1=49$的t分布(参看4.3.2节). 当然实际上不必代入这个公式去手工计算了, 让计算机去代劳好了, 使用R代码

```
x=scan("sugar.txt")#读入数据
t.test(x,m=500,alternative="less")
```

计算结果是$t = -2.6962$(也称为t值), 同时得到p值为0.004793. 看来可以选择显著性水平为0.005, 并宣称拒绝零假设, 而错误拒绝的概率为0.005. 对于这里红糖的具体问题则可以认为, 红糖平均重量为包装上标记的500g是不能接受的, 该数据倾向于支持平均重量少于500g的备选假设. 图6.2给出一个t分布密度函数图, 显示出到底这个t统计量取值在什么位置. 看得出来, 在直观上这也的确是个小概率事件.

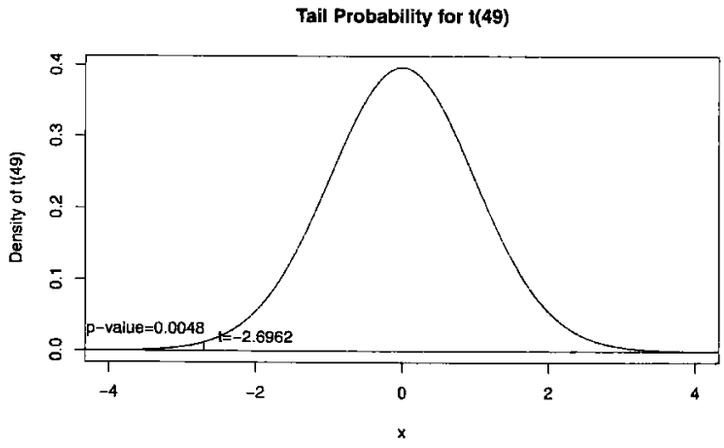

图 6.2 统计量$t = -2.696$相应于左边尾概率(p值)0.0048.

上面例子的备选假设为小于("<")某个值. 同样也可能有备选假设为均值大于(">")某个值的情况. 这种取备选假设为均值大于或小于某个值的检验称为单

尾检验(one-tailed test, 也称为**单侧检验**或**单边检验**). 下面看一个选假设为均值大于(“>”)某个值的例子.

例6.2(数据: exh.txt) 汽车厂商声称其发动机排放标准的一个指标平均低于20个单位. 在抽查了10台发动机之后, 得到下面的排放数据:

17.0 21.7 17.9 22.9 20.7 22.4 17.3 21.8 24.2 25.4

该样本均值为21.13. 究竟能否由此认为该指标均值超过20? 这次的假设检验问题就是

$$H_0:\ \mu = 20 \ \Leftrightarrow\ H_1:\ \mu > 20.$$

和前面的例子的方法类似, 使用R代码

```
x=scan("exh.txt")#读入数据
t.test(x,m=20,alternative="greater")
```

计算结果是$t = 1.2336$, 同时得到p值为0.1243. 这个p值较大, 因此, 没有证据否定零假设. 也可以画出类似于图6.2的尾概率图(图6.3)这时的t分布的自由度为9.

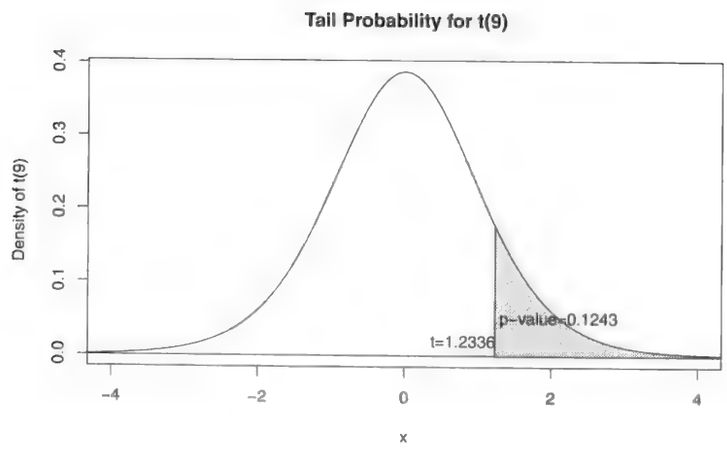

图 6.3　统计量$t = 1.2336$相应于右边尾概率(p值)0.1243.

从图6.3可以看出, 右边的尾概率不能说是小概率. 如果要是拒绝零假设的话, 犯错误的概率就多于12%(0.1243)了, 因此没有足够证据来拒绝零假设, 或者说该检验不显著.

注意: 在假设检验中往往也用带等号的不等式来表示零假设, 比如上述的检验可记为

$$H_0:\ \mu \le 20 \ \Leftrightarrow\ H_1:\ \mu > 20$$

但这里用于计算p值的零假设还是$\mu = 20$, 显然, 如果能够拒绝零假设$\mu = 20$, 那么对于任何μ小于20的零假设就更有理由拒绝了. 这和以拒绝零假设为初衷的假

设检验思维方式是一致的. 在这种记号下, 在不能拒绝零假设时, 如果用"接受零假设"的说法, 就更显得不妥了.

另外, 还有所谓的**双尾检验(two tailed test, 也称为双侧检验或双边检验)**问题, 即

$$H_0: \ \mu = \mu_0 \ \Leftrightarrow \ H_1: \ \mu \neq \mu_0$$

的问题. 显然, 在这种情况下, 尾概率不仅是左边或右边的一个尾概率, 而是两边尾概率之和. 因此如果是一个单尾检验问题, 用了双尾检验的模式, p值就比用单尾检验时大了一倍. 如果在上面例6.2中, 把发动机排放指标例子的检验问题改为是否该发动机的排放指标均值等于20. 检验问题则可以写成

$$H_0: \ \mu \leq 20 \ \Leftrightarrow \ H_1: \ \mu \neq 20$$

这时t统计量还是取原来的值1.2336, 但p值为$0.1243 \times 2 = 0.2486$. 图6.3就变成图6.4的样子.

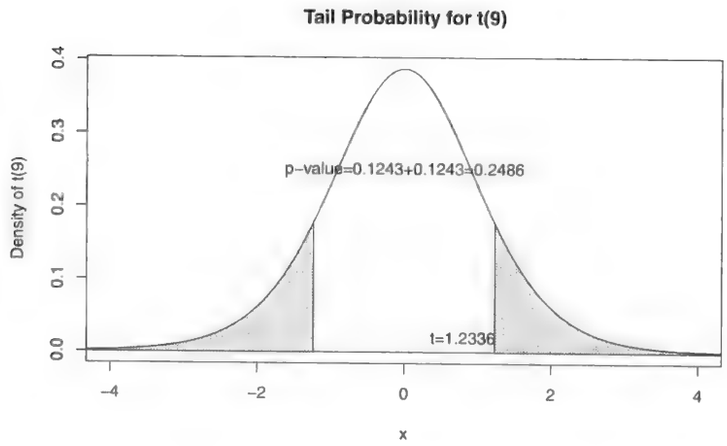

图 6.4 统计量$t = 1.2336$, 而p值等于左右尾概率之和, 即0.2486.

这看起来有些怪异. 好端端的单尾检验为什么要用双尾检验? 对于这个例子, 的确没有必要进行双尾检验.

6.2.2 根据来自两个总体的独立样本对其总体均值的检验

和区间估计类似, 也可以做关于两个独立正态总体均值μ_1和μ_2的差异的假设检验. 和一个总体均值的检验类似, 检验统计量也有t分布. 也可以做单尾和双尾检验. 现用下面例子说明.

例6.3 (数据: drug.txt) 为检测某种药物对攻击性情绪的影响, 对处理组的100名服药者和对照组的150名非服药者进行心理测试, 得到相应的某指标. 人们要检验处理组指标的总体均值μ_1是否大于对照组的指标的总体均值μ_2. 这里, 假定两个总体独立地服从正态分布. 相应的假设检验问题为:

$$H_0: \ \mu_1 = \mu_2 \ \Leftrightarrow \ H_1: \ \mu_1 > \mu_2,$$

这也可以写成

$$H_0:\ \mu_1-\mu_2=0\ \Leftrightarrow\ H_1:\ \mu_1-\mu_2>0.$$

数据有两个变量, 一个是心理测试变量名ah, 另一个是区分两组人的变量id(1为服药组, 2为对照组). 使用R代码

```
w=read.table("drug.txt",header=T)#读入数据
x=w[w[,2]==1,1];y=w[w[,2]==2,1]   #分开两个数据
t.test(x,y,alt="greater")          #检验
```

得到t统计量等于0.9419, p值为0.1736. 因此无法拒绝零假设, 即不能得出处理组的平均指标大于对照组的结论.

说明: 与5.3.2节的关于$\mu_1-\mu_2$的置信区间问题一样, 很多经典文献也建议先做方差是否相等的检验(可用R代码`var.test(x,y)`实行, 得到相应的p值, 对本例, 该检验p值为0.07327091), 如果p值较大, 则用方差相等的公式(相应于R代码`t.test(x,y,alt="greater",var=T)`), 否则用复杂公式(相应于R代码`t.test(x,y,alt="greater")`). 这是前计算机时代节省计算量的产物. 实际上, 任何两个总体的方差都不可能完全相同, 如果相信数据, 不去检验方差, 直接用方差不等的方法去做, 不会有问题的, 即使方差相等, 结果差别也不大.

6.2.3 成对样本的问题

经常有所谓**成对样本(paired samples)**问题. 下面看一个例子.

例6.4(数据: diet.txt) 这里有两列50对减肥数据. 其中一列数据(变量是before)是减肥前的重量, 另一列(变量是after)是减肥后的重量(单位: 公斤). 人们希望比较50个人在减肥前和减肥后的重量. 这样就有了两个样本, 每个都有50个数目. 这里不能用前面的独立样本均值差的检验, 这是因为两个样本并不独立. 每一个人减肥后的重量都和自己减肥前的重量有关, 但不同人之间却是独立的. 令所有个体减肥前后重量差(减肥前重量减去减肥后重量)的均值为μ_D, 这样所要进行的检验为

$$H_0:\ \mu_D=0\ \Leftrightarrow\ H_1:\ \mu_D>0.$$

人们可以把两个样本中配对的观测值逐个相减, 形成由独立观测值组成的一个样本, 然后用单样本检验方法, 看其均值是否为零. 在相减之后公式和单样本均值检验无异. 当然, 如果直接选用软件中成对样本均值的检验, 就不用事先逐个相减了. 这里也有单尾和双尾检验. 这里用的检验是假定总体分布为正态分布时的t检验. 根据R代码

```
w=read.table("diet.txt",header=T)#读入数据
t.test(w$before, w$after, alt="greater",pair=T)#直接检验
```

或者

```
t.test(w$before-w$after, alt="greater") #相减后检验
```

都得到检验统计量$t = 3.355$, p值为0.0007694. 因此在显著性水平为0.001(甚至0.0008)时, 可以拒绝零假设. 也就是说, 减肥后和减肥前相比, 平均重量显著要轻一些.

6.2.4 关于正态性检验的问题

1. 这里对于总体均值的检验均假定了总体分布的正态性, 但如何检验正态性呢(也只能是拒绝或不拒绝)? 最简单实用的方法是用Shapiro正态性检验(Shapiro-Wilk normality test). 它检验:

$$H_0: \text{数据来自正态总体} \Leftrightarrow H_1: \text{数据不是来自正态总体}.$$

比如, 对于sugar数据, 在R中读入数据: `x=scan("sugar.txt")`, 用语句`shapiro.test(x)`, 得到p值为0.4236, 因此没有证据拒绝该变量的正态性. Shapiro检验是一个比较好的检验, 在检验正态性方面一般要优于Kolmogorov-Smirnov检验.

2. 关于检测正态性的直观办法为正态QQ图(不一定准确), 它是用样本分位数与正态分位数做散点图, 对于sugar数据的样本(如果存在变量x中), 在R中可以用下面语句实现: `qqnorm(x);qqline(x)`(图6.5). 如果总体是正态的, 则图上的点应该近似地排成一条直线.

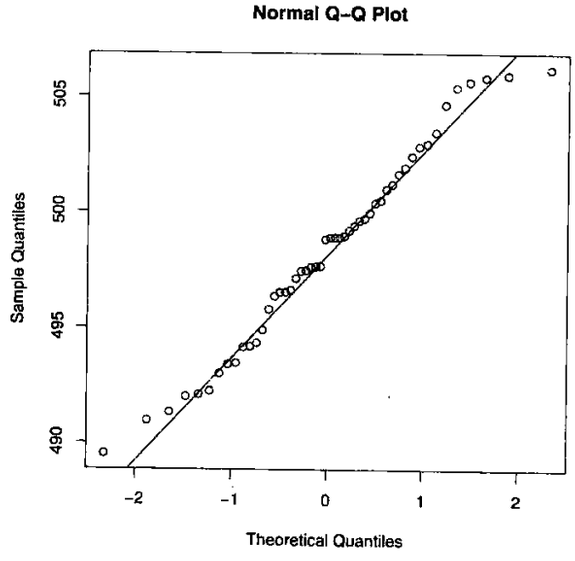

图 6.5 Sugar数据的正态QQ图.

3. 如果正态性假定被拒绝了. 那么这里6.2节的所有检验就都不适用了, 这时可试试后面介绍的非参数检验方法. 那里对总体的假定很少或者没有.

4. 后面在回归和其他一些问题中, 也需要一些正态假定, 也可以使用Shapiro检验来判断假定的合理性.

5. 和任何检验一样, Shapiro检验不能由于p值大就证明一个数据的背景分布为正态. 读者可以试试`shapiro.test(1:30)`, `shapiro.test(1:50)`, `shapiro.test(1:53)`等语句. 实际上, 正整数序列完全不是正态的, 但一直到从1开始的52个整数时, p值才小于0.05.

6.3 对于比例的检验

6.3.1 对于总体比例的检验

例6.5(数据: twop.txt) 对于电视节目, 收视率是个重要的指标. 一个对1500人的电话调查表明, 在某一节目播出的时候, 被访的正在观看电视的人中有23%的正在观看这个节目. 现在想知道, 这是否和该节目的制作人所期望的$p_0 = 25\%$有显著不足.

这个例子可以看成是一个参数为p的二项分布问题(请不要把这个p和检验中的p值混淆!). 形式上的假设检验问题是

$$H_0:\ p = 0.25 \ \Leftrightarrow\ H_1:\ p < 0.25.$$

如果n为访问的正在看电视的人数, x为其中观看该节目的人数, 那么样本中的观看比例为$\hat{p} = x/n = 0.23$. 这是个二项分布的问题, 只要求出在零假设为二项分布$Bin(n, 0.25)$时, 概率$P(x < 0.23n)$就得到p值(用R语句`pbinom(0.23*1500,1500,.25)`得到: 在$n = 1500$时, p值为0.0384). 或者直接用R精确检验语句`binom.test(0.23*1500,1500,.25,alt="less")`得到同样结果.

历史上的近似方法. 在n很大时, 可以用大样本正态近似[1], 那时检验统计量则是在零假设下当大样本时近似有标准正态分布的统计量

$$z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} = \frac{0.23 - 0.25}{\sqrt{\dfrac{0.25(1-0.25)}{n}}}.$$

这个数值用手算也不费力气. 实际上, 对于$n = 1500$, 前面用过的R函数`prop.test()`就是基于这个公式(在做了连续性修正之后[2]) 算出p值为0.03929, 而不做连续性修正直接用公式得到的p值为0.03681914.

给出样本量的情况. 请注意, 前面第五章提起过, 即使被访者的百分比$\hat{p}$都一样, 但样本少的信息就少. 对于假设检验也是一样. 样本量对于假设检验的结果就十分重要. 对于本例, 如果只知道百分比, 下面看看不同的样本量会得到什么结果(假定$\hat{p} = x/n = 0.23$不变).

---

[1] 已经得到精确检验结果了, 为什么还要讲大样本近似检验呢? 这是因为多数传统统计教科书还是习惯于介绍前计算机时代的数学结果, 这里为了尊重统计发展的历史, 也介绍有关的公式和方法.

[2] 连续性修正是指在用连续分布近似离散分布时所做的修正. 例如, 对于取整数值的离散变量的概率$P(2 < X \le 3)$在连续分布下换成$P(2.5 < X < 3.5)$. 而且, 连续性修正可能会有多种模式.

1. 假定样本量为$n = 1500$(和数据给的一样), 已经知道精确检验得到的p值为0.0384, 而(连续性修正后的)正态近似的p值为0.03929. 因此, 可以认为(如果选显著性水平为0.05的话)说收视率有25%是过分了, 即拒绝零假设.

2. 假定样本量为$n = 100$, 那么, 上面的检验通过计算得到的精确p值为0.371(用语句`pbinom(0.23*100,100,.25)`), 而(连续性修正后的) 正态近似的检验得到的p值为0.3645(R语句`prop.test(0.23*100,100,.25,alt="less")`). 因此, 没有足够的理由拒绝收视率有25%的零假设(如果选显著性水平为0.05的话).

读者已经注意到了, 精确检验、利用公式的近似检验以及用连续性修正的近似检验的三种结果都不一样. 在计算机软件很方便的今天, 当然尽量用精确检验了, 而软件通常会自动在样本量太大时自动转换成使用某种连续性修正的近似(不仅仅对正态近似)检验. 代近似公式计算是计算机不发达时的遗产.

前面对总体比例的检验所用的公式利用了二项分布的大样本正态近似, 怎样才能够算是大样本呢? 这里只给出一个必要条件, 这和第五章求比例的置信区间时大样本的近似标准类似, 即当区间

$$p_0 \pm 3\sqrt{\frac{p_0(1-p_0)}{n}}$$

完全包含在(0, 1)区间内部时, 一般就近似地认为样本足够大, 能够用正态近似. 另外, 关于比例的检验除了例子中的左边单尾检验之外, 还有右边的单尾检验和双尾检验. 这与均值的检验类似. 详情请看后面的公式.

对于两个样本, 也有关于两个总体比例之差$p_1 - p_2$的检验. 还拿收视率为例, 如果节目甲的样本收视率为20%($\hat{p}_1 = x_1/n_1 = 0.20$), 而节目乙的收视率为21%($\hat{p}_2 = x_2/n_2 = 0.21$), 是不是节目甲的总体收视率就真的低于节目乙? 这就是检验问题

$$H_0:\ p_1 - p_2 = D_0 = 0 \ \Leftrightarrow\ H_1:\ p_1 - p_2 < 0.$$

这里的零假设意味着节目甲和节目乙收视率相等. 检验统计量同样不复杂. 假定$n_1 = 1200, n_2 = 1300$. 使用精确检验的R语句为

```
binom.test(c(.2*1200,.21*1300),c(1200,1300),alt="less")
```

得到p值为0.07882. 这说明对于显著性水平$\alpha = 0.05$, 没有足够证据拒绝零假设.

历史上的近似方法. 传统的教科书都表明, 该检验统计量在零假设下在大样本时为具有近似标准正态分布的统计量

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} = \frac{(0.20 - 0.21) - 0}{\sqrt{\dfrac{0.2(1-0.2)}{1200} + \dfrac{0.21(1-0.21)}{1300}}}.$$

根据这个公式, "手算"也可以得到结果, 由此得到p值等于0.2679. 因此, 在显著性水平即使是0.1时, 也没有足够证据可以拒绝"节目甲和节目乙收视率相等"的零假设. 用使用连续性修正的R语句

```
prop.test(c(.2*1200,.21*1300),c(1200,1300),alt="less")
```

得到检验的p值为0.2847. 这个结果和不用连续性修正的差不多, 但和精确检验的0.07882还是很不同. 这再次表明, 传统的套用近似数学公式的算法, 即使有计算机软件, 也最好不用, 能用精确检验就一定首先使用.

6.3.2 对于连续变量比例的检验

有时需要检验收入低于某个水平的人占有的比例p是否和预期的p_0一样. 这里的原理和6.3.1节一样, 只要把大于某水平的观测值看作Bernoulli试验的"成功", 而把小于某水平的观测值看成"失败", 就回到二项分布的问题了. 当然, 用不着把这些连续变量的观测值都变成"成功"和"失败"之后, 再数各有多少. 统计软件会替你做所有的事情. 下面通过一个例子来说明.

例6.6 某微生物的寿命问题(数据: life.txt) 这里有某微生物在一种污染环境下生存的寿命数据(单位: 小时)

```
2.12 2.21 0.26 0.04 0.27 2.12 0.42 1.73 0.22 0.12 1.89 0.72 1.73 0.26 1.99
0.28 0.87 0.33 5.65 0.18 1.60 3.13 0.96 0.73 0.74 1.52 0.18 0.85 1.49 0.31
1.01 1.51 1.79 2.30 0.02 0.06 1.44 0.08 0.14 1.14 0.01 0.15 1.50 0.14 0.69
3.45 1.05 0.47 0.06 0.16 0.13 1.15 0.04 1.89 0.34 0.08 0.24 2.64 0.95 0.14
```

问题是存活时间低于2小时的是否少于70%(存活时间多于2小时的是否多于30%)? 因此, 问题的零假设为存活时间低于2小时的少于或等于70%, 而备选假设为存活时间低于2小时的多于70%. 该检验用数学语言表示为, 对于$p_0 = 0.7$,

$$H_0:\ p = p_0 \ \Leftrightarrow\ H_1:\ p > p_0.$$

使用R语句

```
x=scan("life.txt") #读入数据
binom.test(sum(x<2),60,.7,alter="greater") #检验
```

可得p值等于0.002208, 而且还表明活不到两小时的有52个. 因此, 可以拒绝"存活时间低于2小时的少于70%"的零假设.

这个检验的假设还可以有另一种等价形式. 前面第三、四章介绍过样本和总体的α分位数的概念. 例6.6的检验问题等价于检验0.7分位数q是等于2($q_0 = 2$) 还是小于2, 即检验: $H_0:\ q = q_0 = 2 \ \Leftrightarrow\ H_1:\ q < q_0 = 2$. 该例的结论是实际存活时间的0.7分位数$q$小于2小时.

上面的检验又称为(推广的)**符号检验(sign test)**. 它用不着对总体分布进行任何假定. 而狭义的符号检验是指上面的$p_0 = 0.5$或者(等价地) q_0等于中位数的情况. 通常把符号检验归于非参数检验范畴(见下一节).

6.4 非参数检验

6.4.1 关于非参数检验的一些常识

什么是非参数检验?

前面很多检验都假定了总体的背景分布. 但也有些检验没有假定总体分布的具体形式, 这些检验多根据数据观测值的相对大小建立检验统计量, 然后找到在零假设下这些统计量的分布, 并且看这些统计量的数据实现是否在零假设下属于小概率事件. 这种和数据本身的总体分布无关的检验称为非参数检验. 除了本节会介绍一些非参数检验之外, 本书其他章节也有一些非参数检验的例子: 比如, 前面对于连续变量比例的检验, 第七章列联表中的Fisher精确检验, 列联表分析中的Pearson χ^2检验和似然比χ^2检验等都可以认为属于非参数检验范畴.

非参数检验有什么优越性?

非参数检验在总体分布未知时有很大的优越性. 在分布未知时, 如果还假定总体有诸如正态分布那样的已知分布, 在进行统计推断就可能产生错误甚至灾难. 非参数检验总是比传统检验安全. 但是在总体分布形式已知时, 非参数检验就不如传统方法效率高. 这是因为非参数方法利用的信息要少些. 往往在传统方法可以拒绝零假设的情况, 非参数检验无法拒绝. 用统计的术语来说, 在总体分布已知时, 传统方法有较大的势(power), 效率要高, 但非参数统计在总体分布未知时效率要比假定了错误总体分布时的传统方法要高, 有时要高很多.

如何比较检验的效率?

这里所说的效率通常用两种检验方法的**渐近相对效率(ARE)**来度量. 当ARE等于1时表示两者效率一样. 用后面要介绍的Wilcoxon检验为例, 在与通常的t检验比较时, 如果已知总体是正态分布, Wilcoxon检验相对于t检验的ARE为0.864, 而当总体未知时, 它相对于t检验的ARE在某些情况可以达到无穷大. 由此可见非参数检验的优点. 是否用非参数统计方法, 要根据对总体分布的了解程度来确定. 一般来说, 检验T_1对检验T_2的相对效率是这两个检验在拒绝零假设时所使用的最小样本量n_1和n_2的反比: n_2/n_1. 显然, 用的样本量越少, 效率越高. 渐近相对效率是在固定显著性水平α时, 保持两个检验的势(即不犯第二类错误的概率: $1-\beta$)不变时, 让$n_1 \to \infty$, 这时, 为了保持势一样, 必然也有$n_2 \to \infty$, 而n_2/n_1的极限就叫做检验T_1对检验T_2的渐近相对效率.

秩的概念

非参数检验中**秩(rank)**是最常使用的概念. 什么是一个数据的秩呢? 一般来说, 秩就是该数据按照升幂排列之后, 每个观测值的位置. 以下面数据为例(样本量为10):

15　9　18　3　17　8　5　13　7　19

该数据按照升幂重新排列, 成为

3　5　7　8　9　13　15　17　18　19

加上它们的大小次序号(这就是它们的秩), 得到

观测值	3	5	7	8	9	13	15	17	18	19
秩	1	2	3	4	5	6	7	8	9	10

这样, 按照原先的数据次序就是

X_i	15	9	18	3	17	8	5	13	7	19
R_i	7	5	9	1	8	4	2	6	3	10

这下面一行(记为R_i)就是上面一行数据X_i的秩. 如果数字有重复, 那么会有两个秩一样, 比如1, 2, 2, 3四个数字的秩为1, 2.5, 2.5, 4.

利用秩的大小进行推断就避免了不知道背景分布的困难. 这也是大多数非参数检验的优点. 多数非参数检验明显地或隐含地利用了秩的性质, 但也有一些非参数方法没有涉及秩的性质.

一些非参数检验的计算往往有多种选择. 比如列联表分析中的许多问题都有精确方法、Monte Carlo抽样方法和用于大样本的渐近方法等选择. 精确方法比较费时间, 后两种要粗糙一些, 但要快些.

思考一下:

1. 传统的检验是在产生数据的总体分布已知时所用的. 实际上, 人们没有任何办法来证明一个数据产生于某确定总体. 因此, 最多只能说, 用某个检验没有发现足够证据来否认一个数据来自某总体.

2. 除了本节所涉及的基于秩的非参数统计内容之外, 非参数统计还有另外一个领域, 即非参数密度估计和非参数回归等. 它和基于秩的非参数统计从目的到方法很不一样.

6.4.2　关于单样本位置的符号检验

前面介绍过关于位置参数均值的t检验. 那里需要假定观测值的总体分布是正态分布. 如果人们对总体分布一无所知, 就不能假定正态分布, 也不能进行t检验了. 这时, 就可以用**符号检验(sign test)**, 它是对位置参数中位数的检验, 而且不需要任何关于总体的假定. 当然, 对于像正态分布或t分布那样的对称分布, 总体中位数就是总体均值, 这时, 对中位数的检验等价于对均值的检验.

其实, 前面已经通过例6.6介绍过符号检验(sign test)以及推广的符号检验了. 那里的检验是以两种等价的形式出现的, 一种是看中位数或α分位数是否是某个事先认定的值(零假设), 一种是大于(或小于)某数的观测值是否为一个事先认定的比例(零假设).

由于在6.3.2节已经对广义的符号检验进行了较详尽的分析. 这里仅仅通过一个例子对于较简单的狭义符号检验作一描述. 也当成是对6.3.2节的复习吧.

例6.7 西洋参数据(gs.txt) 质量监督部门对商店里面出售的某厂家的西洋参片进行了抽查. 对于25包写明为净重100g的西洋参片的称重结果为(单位: 克):

99.05	100.25	102.56	99.15	104.89	101.86	96.37	96.79	99.37
96.90	93.94	92.97	108.28	96.86	93.94	98.27	98.36	100.81
92.99	103.72	90.66	98.24	97.87	99.21	101.79		

用m表示总体中位数. 容易计算出, 样本中位数为98.36. 因此, 人们怀疑厂家包装的西洋参片份量不足. 由于对于这些重量的总体分布不清楚, 决定对其进行符号检验. 需要检验的是:

$$H_0 : m = 100 \Leftrightarrow H_a : m < 100.$$

按照零假设, 每个观测值(每包西洋参的净重)大于中位数$m_0 = 100$g的机会和小于100g的概率都是0.5. 这服从二项分布$Bin(25, 0.5)$. 容易算出, 大于100g的只有8包. 这样, 参数为$n = 25, p = 0.5$的二项分布变量小于或等于8的概率为0.05388. 这就是p值. 因此, 对于显著性水平$\alpha = 0.05$, 根据这个符号检验, 没有充分的证据拒绝零假设. 这个计算的代码(包括读入数据)为:

```
x=scan("gs.txt");pbinom(sum(x>100),25,.5)
```

大于零假设中位数m_0的个数等于所有观测值减去m_0之后所得的符号为正的差的个数, 而小于m_0的个数等于符号为负的差个数. 上面例子中正号的有8个(用语句`sum(x>100)`), 负号的有17个(用语句`sum(x<100)`). 这就是这个检验之所以被称为符号检验的原因. 本例中没有等于100克的包装. 如果有等于100的, 则既不相应于属于正号, 又不相应于负号, 对判断没有贡献, 一般就把它删除了. 但对于连续型变量, 只要不过分四舍五入, 不大可能出现刚好等于某预先确定值的情况.

6.4.3 关于单样本位置的Wilcoxon符号秩检验

前面介绍的符号检验利用了观察值和零假设的中位数之差的符号来进行检验, 但是它并没有利用这些差的绝对值的大小所包含的信息. 不同的符号仅仅代表了在中位数的哪一边, 而差的绝对值的秩的大小代表了距离中心的远近. 如果把这二者结合起来, 自然比仅仅利用正负号的数目要更有效. 这也是下面要引进的**Wilcoxon 符号秩检验(Wilcoxon signed-rank test)**的宗旨. 它把差的绝对值的秩分别按照不同的符号相加作为其检验统计量.

注意, 和符号检验不同. Wilcoxon符号秩检验需要一点对数据总体分布的了解或假定, 它要求假定样本点来自连续对称总体分布, 而符号检验不需要知道任何总体分布的性质. 在对称分布中, 总体中位数和总体均值是相等的, 因此, 对于来自连续对称总体的数据来说, 对总体中位数的检验, 等价于对于总体均值的检验.

Wilcoxon符号秩检验的原理是这样的. 假定$x_1, ..., x_n$为来自连续对称总体的一个样本. 如果问题的零假设为中位数(均值)$m = m_0$, 对于符号检验

只要计算在n个差$x_i - m_0$ $(i = 1, ..., n)$中有多少正负符号, 即可利用二项分布的概率来计算p值. 但对于Wilcoxon符号秩检验, 则要把那些$|x_i - m_0|$排序, 得到$|x_i - m_0|$的秩. 然后把$x_i - m_0$的符号加到相应的秩上面. 于是, 可以得到既有带正号的秩, 又有带负号的秩. 对带负号的秩的绝对值求和, 即把满足$x_i - m_0 < 0$的$|x_i - m_0|$的秩求和, 并用W^-表示. 类似地, 对带正号的秩的绝对值也求和, 即把满足$x_i - m_0 > 0$的$|x_i - m_0|$的秩求和, 并用W^+表示[1]. 如果m_0的确是中位数, 那么, W^-和W^+应该大体上差不多. 如果W^-或者W^+过大或过小, 则怀疑中位数$m = m_0$的零假设. 令$W = \min(W^-, W^+)$, 则当W太小时, 应该拒绝零假设. 这个W就是Wilcoxon符号秩检验统计量. 一般的书上都有其分布表. 当然, 用不着查表来得到p值, 计算机会做所有的繁琐事情.

下面用例6.7来说明Wilcoxon符号秩检验. 当然, 应该先假定例6.7的样本来自对称的连续总体分布才行. 这里的检验和前面的符号检验的目的一样, 也是检验

$$H_0 : m = 100 \Leftrightarrow H_a : m < 100.$$

使用R语句`wilcox.test(x,m=100,alt="less")`, 得到Wilcoxon符号秩检验的p值为0.04763. 这比前面的符号检验的p值(0.05388)要小, 所以, 如果数据来自对称分布, 用Wilcoxon符号秩检验比符号检验效率要高, 在显著性水平$\alpha = 0.5$时, 可以拒绝零假设.

6.4.4 关于随机性的游程检验(runs test)

游程检验方法是检验一个取两个值的变量的这两个值的出现是否是随机的. 假定下面是由0和1组成的一个这种变量的样本(数据run1.sav):

0 0 0 0 1 1 1 1 1 1 0 0 1 0 1 1 1 0 0 0 0 0 0 0 0

其中相同的0(或相同的1)在一起称为一个**游程(run)**, 单独的0或1也算. 上面这个数据中有4个由0组成的游程和3个由1组成的游程. 一共是$R = 7$个游程. 其中0的个数为$m = 15$, 而1的个数为$n = 10$. 出现0和1的这样一个过程可以看成是参数为某未知p的Bernoulli试验. 但在零假设: "给定了m和n之后, 0和1的出现是随机的"之下, 游程数目R的条件分布就和这个参数无关了. 根据初等概率论, 在零假设下, R的分布可以写成(令$N = m + n$)

$$P(R = 2k) = \frac{2\binom{m-1}{k-1}\binom{n-1}{k-1}}{\binom{N}{n}}$$

[1]对于假定的连续分布, $x_i - m_0 = 0$的概率为零, 因此不应该出现, 但由于四舍五入的原因, 的确会出现$x_i - m_0 = 0$的情况, 而且在统计量的计算p值时, 也有大量的比较, 可能会出现本应不相等的两个连续变量的值相等的情况, 这时软件会自动转换成近似公式来检验了.

$$P(R=2k+1)=\frac{2\binom{m-1}{k-1}\binom{n-1}{k}+2\binom{m-1}{k}\binom{n-1}{k-1}}{\binom{N}{n}}$$

于是就可以算出在零假设下有关R的概率, 以及进行有关的检验了. 利用上面公式可进行精确检验, 也可以利用大样本的渐近分布和利用Monte Carlo方法进行检验了. 对于数据(run1.txt), 利用程序包tseries[1]中的函数runs.test(), 运行下面R语句(包括读入数据):

```
library(tseries);x=scan("run1.txt");runs.test(factor(x))
```

得到p值为0.01052. 因此对于显著性水平0.05, 可以拒绝零假设, 即认为这个数据的0和1的出现不是随机的.

当然, 游程检验并不仅仅用于只取两个值的变量, 它还可以用于某个连续变量的取值小于某个确定值及大于该值的个数(类似于0和1的个数)是否随机的问题. 看下面例子.

例6.8 化妆品数据(run2.txt) 从某装瓶机出来的30盒化妆品的重量如下(单位: 克)

71.6	71.0	71.8	70.3	70.5	72.9	71.0	71.0	70.1	71.8
71.9	70.3	70.9	69.3	71.2	67.3	67.6	67.7	67.6	68.1
68.0	67.5	69.8	67.5	69.7	70.0	69.1	70.4	71.0	69.9

为了看该装瓶机是否工作正常, 首先需要验证大于和小于中位数的个数是否是随机的(零假设为这种个数的出现是随机的). 如果把小于中位数的记为0, 否则记为1, 上面数据变成下面的0和1的序列

1 1 1 1 1 1 1 1 0 1 1 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0

这就归为上面的问题. 用下面语句

```
library(tseries);x=scan("run2.txt")
runs.test(factor(x>median(x)))
```

得到p值为0.00295. 因此对于大于0.005的显著性水平, 可以拒绝零假设, 即有理由认为这30盒化妆品的重量小于中位数和大于中位数的情况的出现不是随机的. 注意: 这里的R函数runs.test()不是精确检验. 可以很容易地编一个游程检验精确的程序(见后面6.6.2节).

6.4.5 比较两独立总体中位数的Wilcoxon (Mann-Whitney)秩和检验

前面说过的关于正态总体均值的检验需要知道或假定所感兴趣的总体是近似正态分布. 然而, 在许多情况, 这种正态总体的假定是不可靠的. 能否在总体

[1] Adrian Trapletti and Kurt Hornik (2012). tseries: Time Series Analysis and Computational Finance. R package version 0.10-29.

分布不知道的时候有办法来检验两个总体的中位数是否相等呢? 这里介绍常用的**Wilcoxon(或称Mann-Whitney)秩和检验**. 它的原理很简单, 假定第一个样本有m个观测值, 第二个有n个观测值. 把两个样本混合之后把这$m+n$个观测值按照大小次序排序, 然后记下每个观测值在混合排序下面的秩. 之后分别把两个样本所得到的秩相加. 记第一个样本观测值的秩的和为W_X而第二个样本秩的和为W_Y. 这两个值可以互相推算, 称为Wilcoxon统计量. 该统计量的分布和两个总体分布无关. 由此分布可以得到p值. 直观上看, 如果W_X与W_Y之中有一个显著地大(或显著地小), 则可以选择拒绝零假设. 这个检验就称为Wilcoxon秩和检验, 也称Mann-Whitney检验. 之所以有两个名称是因为有两个分别由Wilcoxon和Mann-Whitney导出的检验统计量. 虽然这两个统计量不同, 但它们是等价的. 该检验需要的唯一假定就是**两个总体的分布有类似的形状**(不一定对称).

例6.9 GDP数据(gdp.txt)这是地区1的10个城市和地区2的15个城市城镇人口的人均GDP(元). 现在要想以此作为两个样本来检验两个地区的城镇人口的人均GDP的总体中位数m_1和m_2是否一样, 即双尾检验$H_0: m_1 = m_2 \Leftrightarrow H_a: m_1 \neq m_2$. 由于地区2的样本人均GDP的样本中位数大于地区1的中位数, 因此也可以做单尾检验$H_0: m_1 = m_2 \Leftrightarrow H_a: m_1 < m_2$.

用下面R代码(包括输入数据)做单位检验:

```
w=read.table("gdp.txt")
wilcox.test(w[w[,2]==1,1],w[w[,2]==2,1],alt="less")
```

得到p值为0.008138. 这个例子的结果表明, 如果显著性水平选为0.01, 则可以拒绝原假设, 即有理由认为地区2的人均GDP的中位数要高一些.

6.5 从一个例子说明"接受零假设"的说法不妥

虽然前面已经有了一些例子说明"接受零假设"说法的不妥, 但有些人还可能会对于在检验结果不显著时只能说"不能拒绝零假设"而不能说"接受零假设"感到不解. 下面用一个描述性例子来说明.

例6.10 (数据: rice.txt) 一个大米加工厂卖给一个超市一批标明10kg重的大米. 而该超市怀疑该厂家缺斤短两, 对10包大米进行了称重, 得到下面结果(单位: 千克)

9.93 9.83 9.76 9.95 10.07 9.89 10.03 9.97 9.89 9.87

这里假定打包的大米重量服从正态分布. 由于发生分歧, 于是各方同意用这个数据进行关于大米重量均值μ的检验, 以厂家所说的平均重量为10kg作为零假设, 而以超市怀疑的份量不足10kg作为备选假设:

$$H_0: \ \mu = 10 \ \Leftrightarrow \ H_1: \ \mu < 10.$$

于是, 超市、加工厂老板和该老板的律师都进行了检验. 结果是:

1. **超市用全部数据进行t检验, 得到拒绝零假设的结论.** 他们根据计算(用语句`t.test(x,m=10,alt="less")`)得到: 样本均值为9.92kg, 而p值为0.0106. 因此超市认为, 对于显著性水平$\alpha = 0.05$, 应该拒绝零假设.

2. **大米加工厂老板只用2个数据, 得到"接受零假设"的结论.** 大米加工厂老板也懂些统计, 他只取了上面样本的头两个数目9.93和9.83进行同样的t检验. 通过对这两个数进行计算(用语句`t.test(x[1:2],m=10,alt="less")`)得到: 样本均值为9.88kg, 而p值为0.1257. 虽然样本均值不如超市检验的大, 但p值大大增加. 加工厂老板于是下了结论: 对于水平$\alpha = 0.05$, "接受零假设", 即加工厂的大米平均重量的确为10kg.

3. **大米加工厂老板的律师用了全部数据, 但不同的检验方法, 得到"接受零假设"的结论.** 大米加工厂老板的律师说可以用全部数据. 他利用6.3.2节对于连续变量比例的检验, 也就是关于中位数的符号检验(注意对于正态分布, 对中位数的检验等价于对均值的检验). 根据计算(用语句`pbinom(sum(x>10),length(x),.5)`), 得到该检验的p值为0.0547. 所以这个律师说在显著性水平$\alpha = 0.05$时, 应该"接受零假设". 还说: "既然三个检验中有两个都接受零假设, 就应该接受."

显然后面两个人的做法是不对的, 为什么呢?

加工厂老板实际上减少了作为证据的数据, 因此只能得到"证据不足, 无法拒绝零假设"的结论. 但加工厂老板利用一些统计教科书的错误说法, 把"证据不足以拒绝零假设"说成"接受零假设"了. 而且, 从样本中仅选择某些数目(等于销毁证据)违背统计道德.

律师虽然用了全部数据, 但用了不同的方法. 他也只能够说"在这个检验方法下, 证据不足以拒绝零假设"而不能说"接受零假设". 另外, 律师对超市用更有效的检验方法得到的"拒绝零假设"的结论视而不见, 这也违背了统计原理. 其实, 对于同一个检验问题, 可能有多种检验方法. 但只要有一个拒绝, 就可以拒绝. 那些不能拒绝的检验方法是能力不足. 用统计术语来说, 该拒绝而不能拒绝的检验方法是**势(power)不足**, 或者**效率(efficiency)低**.

关于例6.10的总结

该例进行了对于同样假设检验问题的三次检验, 得到三个结果. 该例说明了几个问题:

1. 在已经得到样本的情况下, 随意舍取一些数目是违背统计原理和统计道德的. 这相当于篡改或毁灭证据.

2. 由于证据不足而不能拒绝零假设绝对不能说成"接受零假设". 如果一定要说, 请给出你接受零假设所可能犯第二类错误的概率(这是无法算出的). 这是加工厂老板和律师所犯的错误.

3. 例中律师的检验和超市所做的检验都针对同样的检验问题, 但由于超市的检验方法比律师的检验更强大(或更强势, more powerful, 更有效率, more efficient), 所以超市拒绝了零假设, 而律师的检验则不能拒绝. 如果有针对同一检验问题的许多检验方法, 那么, 只要有一个拒绝, 就必须拒绝. 绝对不能"少数服从多数", 也不能"视而不见".

4. 以关于均值的t检验为例, 实际上, 只要零假设的均值和样本均值的确不一样, 那么根据检验统计量的公式可以看出, 如果样本量不断增大, 就必然会拒绝零假设. 这从例6.5关于比例的检验也可以看出. 当然, 对于效率较低的检验, 要拒绝零假设所需要的样本量较大.

5. 在本书介绍的各种检验中, 只要样本量充分小, 就必定不能拒绝零假设, 如果这可以解释为"接受零假设"的话, 那么减少样本量就荒谬地成为得到"接受零假设"的最佳途经.

6.6 小结

6.6.1 本章的概括和公式

假设检验是关于总体参数的. 为假设检验所建立的检验统计量的分布是基于零假设的. 备选假设是对立于零假设而立的, 备选假设一般直观上被数据所支持. 最终判断需要看检验统计量所取到的(代入数据所得到的)实现值或更极端(更有利于备选假设)的值的概率而定. 这个概率称为p值. p值越小就越有理由拒绝零假设. 如果零假设为真而拒绝了零假设, 称为犯第一类错误, 如果备选假设为真而接受零假设, 称为犯第二类错误.

1. 假设检验的步骤

第一, 写出零假设和备选假设; 第二, 确定检验方法(前计算机时代要确定检验统计量的公式); 第三, 计算p值, 如果p值小于或等于某头脑中的显著性水平α, 就拒绝零假设, 这时犯错误的概率最多为α, 如果p值大于α, 就不拒绝零假设, 因为证据不足.

在前计算机时代, 上面第三步是确定显著性水平α; 第四步是计算检验统计量的实现值; 第五步为用实现值和表中相应于α的临界值比较来决定检验是否显著.

2. 关于一个正态总体均值的单尾和双尾检验(两个方向的单尾和一种双尾)

$$H_0 : \mu = \mu_0 \Leftrightarrow H_1 : \mu < \mu_0;$$
$$H_0 : \mu = \mu_0 \Leftrightarrow H_1 : \mu > \mu_0;$$
$$H_0 : \mu = \mu_0 \Leftrightarrow H_1 : \mu \neq \mu_0.$$

检验统计量均为在零假设下具有$n-1$个自由度的t分布的统计量

$$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}},$$

据此可以计算尾概率(该统计量取其实现值或更极端值的概率)p值. 如果p值很小, 则拒绝零假设, 否则没有足够理由拒绝.

3. 关于两个独立正态总体均值差的单尾和双尾检验

这里分总体方差相等和不相等两种情况. 实际上, 根本无法根据数据证明两个总体方差相等. 用方差的假设检验可以拒绝方差相等. 但有人觉得可以用证据不足以拒绝方差相等的零假设可以"证明"两个总体方差相等. 这完全是错误的. 因为在小样本时, 基本上都无法拒绝方差相等的假设, 这只能说证据不足, 因而仍然把方差相等作为一个假定(而不是事实!)

$$H_0:\mu_1-\mu_2=D_0 \Leftrightarrow H_1:\mu_1-\mu_2>D_0;$$
$$H_0:\mu_1-\mu_2=D_0 \Leftrightarrow H_1:\mu_1-\mu_2<D_0;$$
$$H_0:\mu_1-\mu_2=D_0 \Leftrightarrow H_1:\mu_1-\mu_2\neq D_0.$$

这里最经常的情况为$D_0=0$的情况.

假定两总体方差相等假定下的公式. 检验统计量为在零假设下具有n_1+n_2-2个自由度的t分布的

$$t=\frac{(\bar{x}_1-\bar{x}_2)-D_0}{\sqrt{s_P^2(\frac{1}{n_1}+\frac{1}{n_2})}}$$

这里

$$s_P^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2},$$

$\bar{x}_1$和$\bar{x}_2$为两个样本的均值, n_1和n_2为两个样本量, 而s_1和s_2为两个样本标准差.

假定两总体方差不相等假定下的公式. 检验统计量为

$$t^*=\frac{(\bar{x}_1-\bar{x}_2)-D_0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}},$$

它近似地具有如下定义的自由度ℓ的t分布:

$$\ell=\frac{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}{\frac{s_1^2}{n_1^2(n_1-1)}+\frac{s_2^2}{n_2^2(n_2-1)}}$$

注意: 在一些教科书中, 关于均值的检验还分大样本和小样本之别, 大样本的统计量近似地用正态分布, 而小样本用t分布. 这是计算机大量使用之前, 完全依赖

查表时的产物. 在计算机软件中, 一般用不着另外设置.

4. **关于一个总体比例的单尾和双尾检验(大样本近似公式)**

$$H_0 : p = p_0 \Leftrightarrow H_1 : p > p_0;$$
$$H_0 : p = p_0 \Leftrightarrow H_1 : p < p_0;$$
$$H_0 : p = p_0 \Leftrightarrow H_1 : p \neq p_0.$$

最简单的办法是在零假设下的二项分布模型$Bin(n, p_0)$下通过概率$P(X < x)$来计算p值. 如非要求近似解, 则可以用在零假设下为近似标准正态的检验统计量(这里记$\hat{p} = x/n$):

$$z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}.$$

而判别大样本的一个粗略的必要条件为

$$p_0 \pm 3\sqrt{\frac{p_0(1-p_0)}{n}}$$

必须完全在(0, 1)区间之内, 否则就说明样本不够大.

5. **关于两个独立总体比例之差的单尾和双尾检验(大样本近似公式)**

$$H_0 : p_1 - p_2 = D_0 \Leftrightarrow H_1 : p_1 - p_2 > D_0;$$
$$H_0 : p_1 - p_2 = D_0 \Leftrightarrow H_1 : p_1 - p_2 < D_0;$$
$$H_0 : p_1 - p_2 = D_0 \Leftrightarrow H_1 : p_1 - p_2 \neq D_0.$$

注意, 上面的D_0在大多数应用情况假设等于0($D_0 = 0$)以检验两个总体比例是否相等. 在零假设下为近似标准正态的检验统计量为(这里记$\hat{p}_1 = x_1/n_1, \hat{p}_2 = x_2/n_2$)

$$z = \frac{\hat{p}_1 - \hat{p}_2 - D_0}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}.$$

6. **关于非参数检验**

非参数检验的精确公式大都没有显式, 或者是一组公式, 或者是一个程序. 但各种大样本近似检验倒是有很多公式, 笔者不想在这里赘述.

6.6.2 R语句的说明

下面仅仅介绍精确检验的算法, 对于前计算机时代的近似算法的遗产, 前面已经介绍得太多, 这里可能仅简单提及.

1. 关于一个正态总体均值的检验

考虑6.2.1节例6.1的红糖例子(数据在sugar.txt), 要检验

$$H_0:\ \mu = 500 \ \Leftrightarrow\ H_1:\ \mu < 500.$$

这是正态总体的均值检验. 可以用x=scan("sugar.txt")来在R中读入数据, 对于一般问题, 根据需要选择下面命令之一:

```
t.test(x,m=m0)                  #(双边检验)
t.test(x,m=m0,alt="greater")  #(右尾检验)
t.test(x,m=m0,alt="less")     #(左尾检验)
```

对于本例, 用t.test(x,m=500,alt="less")即可得到下面输出:

```
        One Sample t-test
data:  x
t = -2.6962, df = 49, p-value = 0.004793
alternative hypothesis: true mean is less than 500
95 percent confidence interval:
     -Inf 499.3749
sample estimates:
mean of x
 498.3472
```

输出中, 除了表明统计量为-2.6962, 自由度为49, p值为0.004793之外, 还说明了备选假设及单边的95%置信区间(检验是单尾的, 置信区间也是半开区间), 最后还给出了样本均值作为均值的估计: 498.3472. 如果只想输出p值, 则用语句t.test(x,m=500,alt="less")$p.value即可(上面输出中的任何值都可以单独输出).

2. 关于两个独立正态总体均值的检验

考虑6.2.2节例6.3的关于服药者的心理测试例子(数据在drug.txt, 这里ah为测试指标, id为区别这两类的代码). 这里用两独立正态总体均值差的检验

$$H_0:\ \mu_1 - \mu_2 = D_0 \ \Leftrightarrow\ H_1:\ \mu_1 - \mu_2 > D_0.$$

在R中, 对于该数据, 在用命令x=read.table("drug.txt",header=T)输入数据之后, 为了符号简单, 我们把两个样本分开:

```
x=w[w[,2]==1,1];y=w[w[,2]==2,1]
```

一个用代码x, 另一个用代码y表示. 对于一般问题, 根据需要选用下面命令之一:

```
t.test(x,y,m=D0)                    #(双边检验)
t.test(x,y,m=D0,alt="greater")  #(右尾检验)
t.test(x,y,m=D0,alt="less")     #(左尾检验)
```

对本例$D_0 = 0$, 用t.test(x,y,alt="greater"), 得到下面输出:

```
        Welch Two Sample t-test
data:  x and y
t = 0.9419, df = 234.348, p-value = 0.1736
alternative hypothesis: true difference
in means is greater than 0
95 percent confidence interval:
 -0.373638       Inf
sample estimates:
mean of x mean of y
    8.598     8.102
```

输出中, 除了表明统计量为0.9419, 自由度为234.348, p值为0.1736之外(这里假定了两总体方差不等), 还说明了备选假设及单边的95%置信区间(也是半开区间), 最后还给出了两个样本均值: 8.598, 8.102. 如果只想输出p值, 则用语句`t.test(x,y,alt="greater")$p.value`即可(上面输出中的任何值都可以单独输出).

3. 成对正态样本的均值检验

使用6.2.3节例6.4减肥数据(diet.txt), 其中一列数据是减肥前的重量, 另一列是减肥后的重量(单位: 公斤). 在R中, 对于该数据, 可以用命令`x=read.table("diet.txt",header=T)`输入数据, 用`attach(x)`之后(把变量名字before和after放入内存), 类似于前面, 也有各种单尾及双尾检验, 对于本数据用`t.test(before, after, alt="greater",pair=T)`或者`t.test(before-after, alt="greater")`可输出同样结果. 结果形式和前面的类似, 这里不再罗列.

4. 关于总体比例的检验

以6.3.1节的例6.5 (数据twop.sav)为例. 由于数据简单, 不用输入数据. 对于单样本检验$H_0: p = 0.25 \Leftrightarrow H_1: p < 0.25$的情况, 如果$n = 1500$, $x = 0.23 \times 1500 = 345$, 则该检验可用`binom.test(345,1500,.25,alt="less")`得到结果, 或者用`pbinom(0.23*1500,1500,.25)`得到p值.

对于后面的两样本问题$H_0: p_1 - p_2 = D_0 = 0 \Leftrightarrow H_1: p_1 - p_2 < 0$, 如果$n_1 = 1200, x_1 = 0.2 \times 1200 = 240$, 而$n_2 = 1300, x_2 = 0.21 \times 1300 = 273$, 则可以用`binom.test(c(240,273),c(1200,1300),alt="less")`得到结果.

5. 关于连续变量比例的检验

以6.3.2节例6.6数据life.txt为例. 在R中, 用`x=scan("life.txt")`输入数据. 用二项分布模型得到精确p值的语句为

```
binom.test(sum(x<2),60,.7,alter="greater")
```

R的综合输出可能会有很多结果同时展示, 但也可以单独输出, 这对编程

很方便. 以6.2.1节例6.1的sugar数据为例, 在用x=scan("sugar.txt")输入数据之后, 如果把检验结果存放到(比如a中):a=t.test(x,m=500,alt="less"), 再用names(a)看有什么可以单独输出的, 这时可以看到屏幕上打印出9个内容:

```
[1] "statistic" "parameter" "p.value" "conf.int" "estimate"
[6] "null.value" "alternative" "method" "data.name"
```

这些内容都可以单独输出. 比如, 可以用a$s (写全了应该是a$statistic)输出t统计量的值, 用a$pa (写全了应该是a$parameter)输出自由度(这里不能简写为a$p, 因为会与a$p.value混淆), 用a$p.v (写全了应该是a$p.value)输出p值等等.

6. **关于总体中位数的符号检验**

就例6.7来说, 只要输入x=scan("gs.txt");pbinom(sum(x>=100),25,.5)即可得出p值.

7. **关于单样本的Wilcoxon符号秩检验**

就例6.7来说, 只要输入wilcox.test(x,m=100,alt="less")即可得出所有结果.

8. **关于随机性的游程检验**

就例6.7来说, 只要输入下面语句

```
library(tseries);x=scan("run2.txt");runs.test(factor(x>median(x)))
```

可得到各种结果, 包括p值等于0.00295. 但这个程序是用的大样本正态近似. 下面给出笔者自己编的精确检验的函数, 它给出了精确p值为0.002261331.

```
runstest=function(y,cut=0){if(cut!=0)x=(y>cut)*1 else x=y
N=length(x);k=1
for(i in 1:(N-1))if (x[i]!=x[i+1])k=k+1; r=k;m=sum(1-x);n=N-m
P1=function(m,n,k)
{2*choose(m-1,k-1)/choose(m+n,n)*choose(n-1,k-1)}
P2=function(m,n,k)
{choose(m-1,k-1)*choose(n-1,k)/choose(m+n,n)+
choose(m-1,k)*choose(n-1,k-1)/choose(m+n,n)}
r2=floor(r/2);if(r2==r/2){pv=0;for(i in 1:r2)pv=pv+P1(m,n,i)
for(i in 1:(r2-1))pv=pv+P2(m,n,i)}else
 {pv=0;for(i in 1:r2)pv=pv+P1(m,n,i); for(i in 1:r2)pv=pv+P2(m,n,i)}
if(r2==r/2)pv1=1-pv+P1(m,n,r2)else pv1=1-pv+P2(m,n,r2)
tpv=min(pv,pv1)*2
list(Exact.pvalue=min(pv,pv1),Exact.2sided.pvalue=tpv)}
```

对于例6.7来说, 只要输入下面语句

```
x=scan("run2.txt"); runstest(x>median(x))
```

就输出了单边和双边的p值. 相信读者能够编出比这个程序更漂亮的程序.

9. **关于比较两个独立总体中位数的Wilcoxon秩和检验**

就例6.9数据而言, 只要输入

```
w=read.table("gdp.txt")
wilcox.test(w[w[,2]==1,1],w[w[,2]==2,1],alt="less")
```

就可以得到结果.

6.7 习题

(1) "假设检验的目的是试图使零假设通过"的说法对吗? 对于本章的例子, 这一点能够做到吗? 举例说明为什么"不能拒绝零假设"并不等于"接受零假设".

(2) 假定有两个班级, 从班级A抽取10个成绩, 它们是6个100分4个99分. 而从班级B抽取两个成绩, 它们是两个负分: -100和-200 (数据: grade6.txt). 这个问题看上去很荒唐, 就当这两个班的老师都很怪异罢了. 现在分别对这两个班进行假设检验(不做比较均值的检验), 零假设是各自总体的平均分数为100分, 而备选假设为各自总体的平均分数小于100分, 即每个检验都是$H_0 : \mu = 100 \Leftrightarrow H_1 : \mu < 100$. 对其进行单样本单尾t检验. 什么是你们的结论? 如果你们觉得结论有趣, 请同学进行讨论, 说出你们对这个题目从提出到结论的任何可能的看法. 你可以选定这两个检验的显著性水平为$\alpha = 0.05$.

(3) 如果关于两个候选人的民意调查表示候选人A有50%的支持率, 而候选人B有48%的支持率, 那么是不是候选人A在整个选民中的支持率一定大于候选人B呢? 我们还缺乏什么信息? 假定这两个样本量分别为500和1200, 你们的结论是什么? 如果两个样本量均为5000呢?

(4) 为了比较两种鞋底材料, 让20名试验者左右脚穿两种不同材料的鞋, 然后记录下左右脚的磨损度(数据shoes.txt). 这是独立样本问题吗? 如果不是, 是什么问题, 为什么? 利用双尾检验, 看两种材料的耐磨度是否一样. 可选显著性水平$\alpha = 0.05$.

(5) 负责任的态度是, 在作出任何结论时都应该给出你的结论可能犯错误的概率. 在假设检验中, 这一点体现在哪里?

(6) 重复本章所有例子的计算, 能够用多种方法的(比如非参数方法), 尽量用多种方法.

(7) 讨论非参数检验的各个方法在总体分布已知时可能存在的经典方法.

(8) 能否试图就非参数检验的一些方法, 举例说明非参数方法的优点.

第七章　变量之间的关系; 回归和分类

7.1　问题的提出

前面的内容大多只涉及一个变量, 是为进一步引进各种概念做铺垫. 世界上任何事物都是互相联系的, 绝大多数真实数据都包含有许多变量的观测值, 这些变量大都是以各种方式相关联的. 统计的主要内容是研究多个变量之间的关系. 例如, 顾客对商品和服务的反映对于商家是至关重要的, 但是仅仅有满意顾客的比例是不够的, 商家希望了解什么是影响顾客观点的因素, 以及这些因素是如何起作用的. 类似地, 医疗卫生部门不能仅仅知道某流行病的发病率, 而且想知道什么变量影响发病率, 如何影响发病率的. 发现变量之间的统计关系, 并且用总结出来的规律来帮助人们进行决策, 这才是统计实践的最终目的.

一般来说, 统计可以根据目前所拥有的信息(数据)来建立人们所关心的变量和其他有关变量的关系. 这种关系一般称为**模型(model)**. 假如用Y表示感兴趣的变量, 用X表示其他可能与Y有关的变量(X也可能是若干变量组成的向量). 则所需要的是建立一个函数关系$Y = f(X)$. 这里Y称为**因变量或响应变量(dependent variable, response variable)**, 而X称为**自变量**, 也称为**解释变量或协变量(independent variable, explanatory variable, covariate)**. 建立这种关系的过程就叫做**回归(regression)**或者**分类(classification)**. 回归和分类的区别在于因变量的性质. 当因变量为数量变量时, 叫做回归, 而当因变量为定量变量(也称名义变量或分类变量)时叫做分类.

思考一下:

1. 一个模型的存在的首要条件是可以很方便地计算. 因此, 后面马上要介绍的线性回归就是在前计算机时代就已经发展出来的可以用手工计算的统计模型之一. 在前计算机时代, 人们必须对数据做出许多主观假定, 才能够进行对数据做基于模型的计算和判断, 而且也只能处理少量数据.

2. 由于人类能力的局限性, 所有的模型都是近似的. 完全准确的模型是不存在的.

3. 经典的统计模型是可以用数学公式描述出来的, 但是, 人们有理由怀疑这些有限的公式对于描述复杂的自然和社会现象的可靠程度. 随着计算机的发展, 就产生了用计算机算法来确定的基于数据本身而不是数学假定的模型, 模型也就变得越来越复杂, 可处理的数据量也越来越大. 这些模型包括机器学习或数据挖掘领域所使用的众多的模型.

一旦建立了回归模型, 除了对各种变量的关系有了进一步的定量理解之外, 还可以利用该模型(函数、关系式或算法)通过自变量对因变量做**预测(prediction)**. 这里所说的预测, 是基于已知的自变量的值, 通过模型对未知的因变量值进行估计, 它并不一定涉及时间先后的概念, 更不必要有因果关系.

下面先看后面还要讨论的数值例子.

例7.1 (数据: bschool.txt) 这是美国60个著名商学院的数据, 包括的变量有读MBA之前的工资(SalaPreMBA, 单位: 千美元)、读MBA之后的工资(SalaPostMBA, 单位: 千美元)、学费(Tuition, 单位: 千美元)、GMAT分数(GMAT, 这是进商学院之前的考试[1]) 等四个变量. 人们想知道读MBA之后的工资和其余几个变量之间的关系, 能不能建立一个回归模型.

对于这个数据, 首先点出每两个变量之间的散点图(图7.1), 这些散点图是用下面代码(包括读入数据)实现的:

```
w=read.table("bschool.txt",header=TRUE);pairs(w)
```

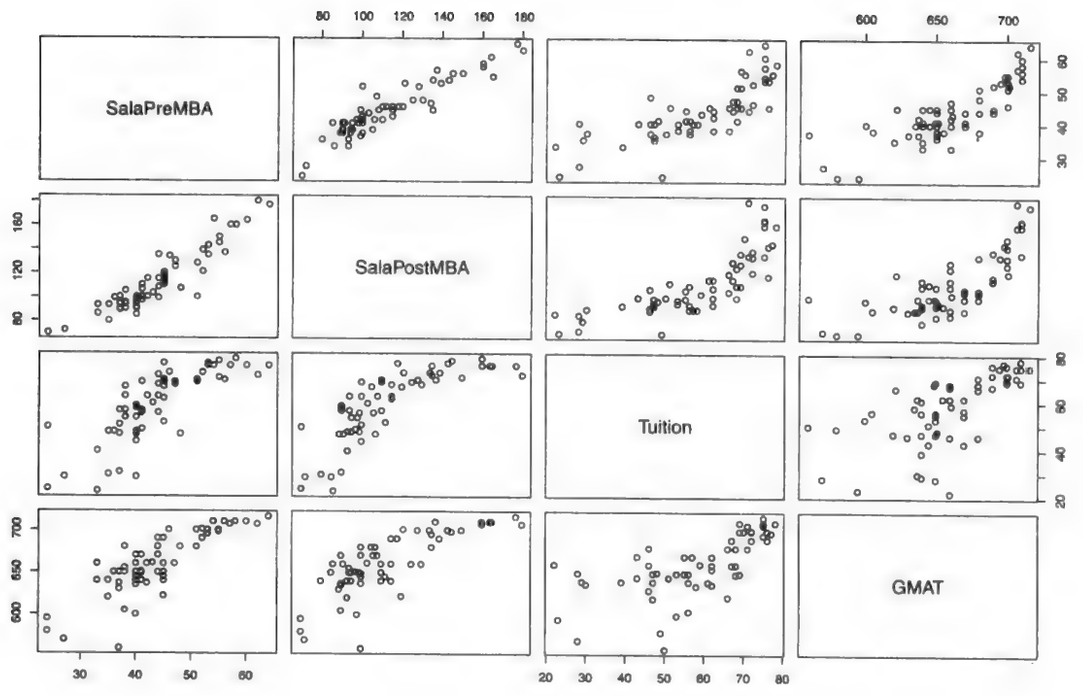

图 7.1　例7.1的60个商学院数据变量的成对散点图.

从图7.1可以看出, SalaPostMBA和SalaPreMBA之间有一个增长, 另一个也增长的某种模式, 而且那些点似乎形成一条直线形状. 其他变量也有类似的模式, 但没有SalaPostMBA和SalaPreMBA之间的关系那么明显. 这种SalaPostMBA和SalaPreMBA之间的那种关系模式可以用**线性相关(linear correlation)** 的术语描述, 下面先介绍线性相关的度量.

7.2　定量变量的线性相关

例7.1中变量之间的线性关系可以用下面几种方法来度量:

[1]GMAT为Graduate Management Admissions Test.

1. **Pearson相关系数(Pearson's correlation coefficient)**, 它又称为**相关系数或线性相关系数**. 它一般用字母r表示. 它是由两个变量的样本$x_1, ..., x_n$及$y_1, ..., y_n$取值得到, 计算公式为

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

这是一个描述线性相关强度的量, 取值于-1和1之间. 当两个变量有很强的线性相关时, 相关系数接近于1(正相关)或-1(负相关), 而当两个变量不那么线性相关时, 相关系数就接近0. 到底r是多少才算是线性相关呢? 有些人给出了下面的粗略的判断方法[1]:

在不同区间的r相对于不同程度的线性相关

相关程度	负	正
不相关	$(-0.09, 0.0)$	$(0.0, 0.09)$
很小	$(-0.3, -0.1)$	$(0.1, 0.3)$
中等	$(-0.5, -0.3)$	$(0.3, 0.5)$
强	$(-1.0, -0.5)$	$(0.5, 1.0)$

相关系数可以用R代码`cor(x,y)`得到. 在数据来自正态总体的假定下, 有相关系数的检验: $H_0 : r = 0 \Leftrightarrow H_1 : r \neq 0$. **要注意的是, 这里仅仅检验是否$r = 0$? 而不是是否线性相关, 因为即使$r \neq 0$也不意味着相关.**

2. **Kendall相关系数(Kendall' s τ)** 这里的度量原理是把所有的样本点配对. 如果每一个点由x和y组成的坐标(x, y)来代表, 一对点就是诸如(x_i, y_i)和(x_j, y_j)那样的点对. 如果样本量为n, 即数据点为$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, 那么一共有$\binom{n}{2} = n!/\{2!(n-2)!\}$这样多的点对. 然后看每一对中的$x$和$y$的观测值是否同时增加(或减少). 比如考虑点对$(x_1, y_1)$和$(x_2, y_2)$, 可以算出乘积$(x_2 - x_1)(y_2 - y_1)$是否大于0, 如果大于0, 则说明$x$和$y$同时增长或同时下降, 称这两点**协同(concordant)**, 否则就是**不协同**. 如果样本中协同的点数目多, 两个变量就更加正相关一些; 如果样本中**不协同(discordant)**的点数目多, 两个变量就更加负相关一些; 如果既不正相关, 也不负相关, 则为不相关. Kendall τ也是在-1和1之间的数, 也是越接近于1或-1就越相关, 而接近0就不相关. 这里不用假设总体的分布, 也可以检验(零假设为$\tau = 0$). 因此Kendall相关系数(记为τ)是一个非参数的度量(所谓非参数方法, 就是它不依赖于变量背后的总体分布).

3. **Spearman 秩相关系数(Spearman rank correlation coefficient 或Spearman' s ρ)** 它和Pearson相关系数定义有些类似, 只不过在定义中把点的坐标换成各自样本的秩(即样本点大小的"座次"). Spearman相关系

[1]不同的人对不同问题的线性相关的理解不一样, 因此判断r到底是多少才算相关, 永远也不可能有一个完全一致的看法.

数(记为ρ)也是取值在-1和1之间, 也有类似的解释. 通过它也可以进行不依赖于总体分布的非参数检验(零假设为$\rho = 0$).

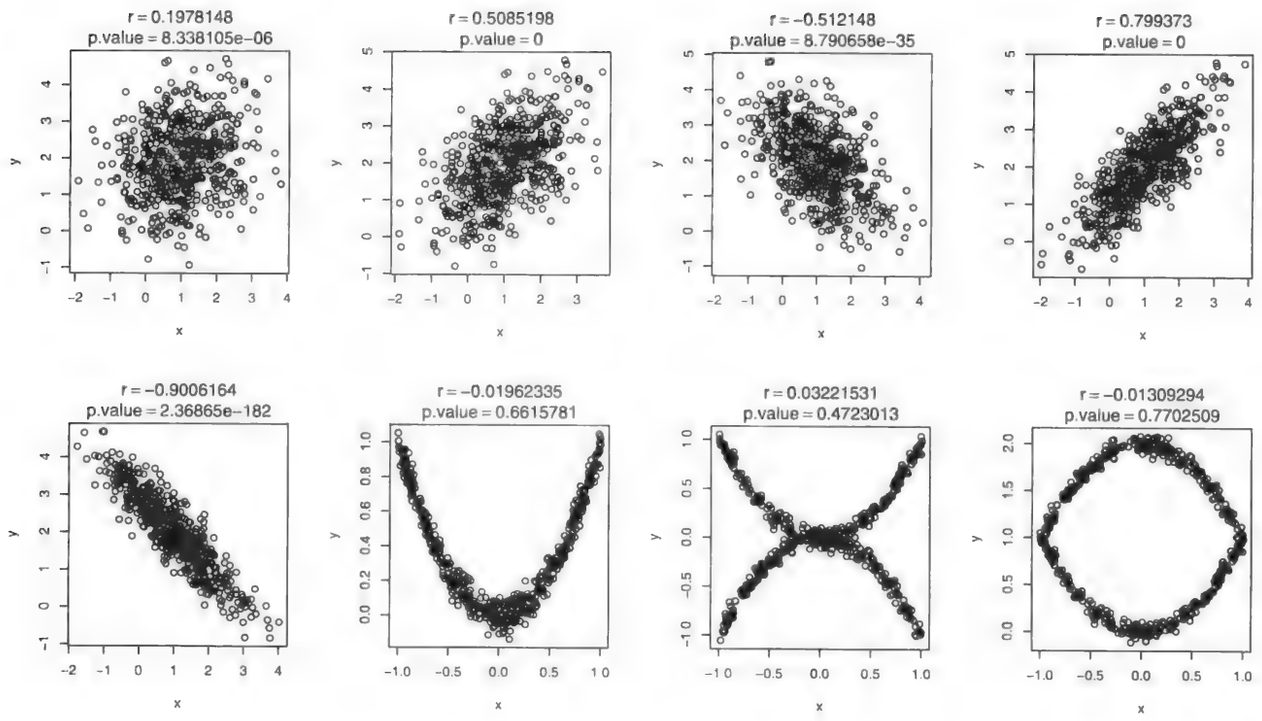

图 7.2　标明相关系数和检验$H_0 : r = 0 \Leftrightarrow H_1 : r \neq 0$的$p$值的8组数据的散点图.

图7.2为8组不同数据的散点图, 图上标明了Pearson线性相关系数及检验$H_0 : r = 0 \Leftrightarrow H_1 : r \neq 0$时的$p$值. 从图7.2可以看出下面几点:

1. 第一行左数第一个图的两组数虽然不线性相关(相关系数不到0.2), 但检验的p值很小(0.000008), 说明这个检验显著(可以拒绝"$r = 0$"的原假设, 但并不相关).

2. 第一行中间两个图(左数第二三个图)看上去仅仅稍微有些相关, 但相关系数的绝对值都在0.5以上(一个正相关, 一个负相关).

3. 第一行最右图和第二行最左图的两组数都线性相关, 前者正相关而后者负相关.

4. 第二行右边三个图中的两个变量显然都很相关, 但不是线性相关, 因此r很小.

再来看例7.1各个变量之间的相关系数, 利用R语句(包括读入数据)

```
w=read.table("bschool.txt",header=TRUE);cor(w)
```

得到下面的线性相关系数表

	SalaPreMBA	SalaPostMBA	Tuition	GMAT
SalaPreMBA	1.000	0.924	0.784	0.825
SalaPostMBA	0.924	1.000	0.781	0.777
Tuition	0.784	0.781	1.000	0.662
GMAT	0.825	0.777	0.662	1.000

看得出来, 这些变量之间相关系数都大于0.6. 下一节我们将考虑以SalaPostMBA为因变量的线性回归模型.

7.3 经典回归和分类

7.3.1 一个数量自变量的线性回归

例7.1 (数据: bschool.txt, 继续) 先考虑一个因变量(SalaPostMBA, 即读MBA前的工资)及一个自变量(SalaPreMBA, 拿到MBA后的工资) 的最简单的情况. 我们在图7.1中看到这两个变量的散点图. 而简单线性回归就是希望能够在图上找到一条直线, 使其能够在某种标准下, 最好地代表这个数据的线性趋势. 当然, 标准不同, 结果也不同. 因此, 首先需要确定选择这条直线的标准. 这里介绍的是**最小二乘回归(least squares regression)**. 古汉语"二乘"是平方的意思. 最小二乘法就是寻找一条直线, 使得所有点到该直线的竖直距离, 即按因变量方向的距离, (该距离称为各个点的**残差, residual**)的平方和最小. 这样的直线很容易通过计算机得到. 这种用模型(这里是一条直线)来近似描述数据的过程也叫做**拟合(fit)**. 这里有两个问题, 一个是为什么考虑竖直距离, 这是因为人们关心对因变量的描述, 自然希望减少在因变量方向的误差. 第二个问题是为什么用残差平方和而不是诸如残差绝对值和等其他度量, 这是因为在前计算机时代, 残差平方和在数学上较易处理, 比如其导数连续等等. 现在已经出现了大量其他的确定回归直线的标准, 各有其优点.

就例7.1数据来着手, 根据计算, 找到进入MBA前的工资(SalaPreMBA)和得到MBA之后的工资(SalaPostMBA)的回归直线. 通过R语句(包括输入数据)

```
a=lm(SalaPostMBA~SalaPreMBA,w);summary(a)
plot(SalaPostMBA~SalaPreMBA,w,pch=16);abline(a)
```

得到下面输出及带有最小二乘回归直线的图7.3.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.4026     6.8394  -1.667    0.101
SalaPreMBA    2.8290     0.1535  18.434   <2e-16
Multiple R-squared: 0.8542,     Adjusted R-squared: 0.8517
F-statistic: 339.8 on 1 and 58 DF,  p-value: < 2.2e-16
```

该输出表明这条直线的截距为-11.4026, 斜率为2.8290. 该直线方程为

$$y = -11.40 + 2.83x$$

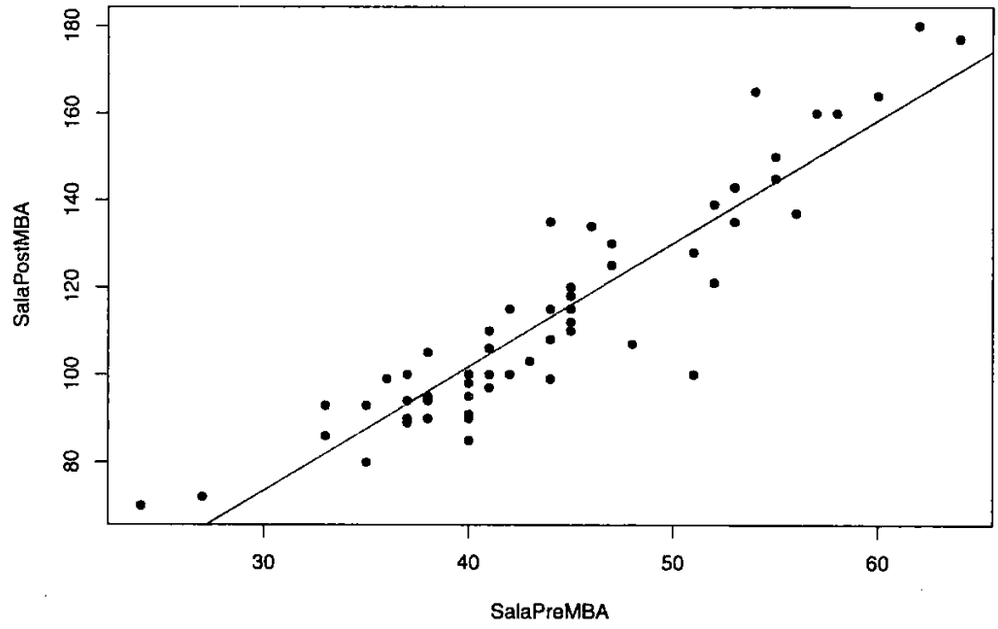

图 7.3　例7.1的SalaPreMBA和SalaPostMBA的散点图及最小二乘回归直线.

这个直线实际上是对所假设的下面线性回归模型的估计, 这里用y代表因变量(这里是SalaPostMBA), x代表自变量(这里是SalaPreMBA).

$$y = \beta_0 + \beta_1 x + \epsilon.$$

这里的ϵ是误差项. 该模型假定, 变量x和y有上面的线性关系, 但凡是不能被该线性关系描述的y的变化都由这个误差项来承担. 由于误差, 观测值不可能刚好在这条直线上, 如果这个模型有道理的话, 这些观测值都不会离这条直线太远. 这里得到的截距和斜率是对β_0和β_1的估计的一个实现, 通常用记号$\hat{\beta}_0$和$\hat{\beta}_1$来记β_0和β_1的估计. 由于不同的样本产生不同的估计, 所以估计量是个随机变量, 也有分布, 也可以用由他们构造检验统计量来检验β_0和β_1是不是显著. 拿回归主要关心的描述两个变量之间关系的斜率β_1来说, 假设检验问题是

$$H_0 : \beta_1 = 0 \Leftrightarrow H_1 : \beta_1 \neq 0.$$

这是个t检验, 如果显著(即可以拒绝零假设), 则认为回归有意义, 也就是说, x的变化与y的变化的确有关. 上面的R输出也给出了这个检验的结果: t检验统计量为18.434, 而p值为2×10^{-16}, 所以该检验很显著. 对于这个数据的残差`a$res`进行了Shapiro-Wilk正态性检验(用R语句`shapiro.test(a$res)`), p值为0.5226, 没有证据拒绝正态性的零假设, 因此不妨假定变量的正态性, 没有正态性的假定, 这个关于系数β_1的检验就值得怀疑了, 还需要满足一些其他条件, 下面予以介绍.

回归中假设检验所需要的条件: 任何回归本身并不需要什么假定的条件, 用手任意画出一条你觉得合适的直线也可以说是回归. 但是要对最小二乘回归系数进行t检验或后面要介绍的关于拟合好坏的F检验就需要对模型作出一些假定. 这

些假定是关于误差项的:

1. ϵ为均值为零的随机变量;

2. ϵ的方差(或标准差)对于所有x的值不变;

3. ϵ互相独立;

4. ϵ为正态分布随机变量.

或者用一句话来叙述这个假定: "ϵ为独立同正态分布的随机变量." 后面要引进的各种回归中的t检验和F检验均需要这些条件. 在前三个条件成立时, 回归模型的误差项ϵ被认为是随机误差, 也就是说, 回归模型是适当的. 但是这些条件并不是自动成立的. 在模型不适当时误差项是不会满足头三个条件的.

除了对β_1的检验之外, 还有一个说明自变量解释因变量变化百分比的度量, 叫做**决定系数(coefficient of determination, 也叫测定系数或可决系数)**, 用R^2(R-squared)表示. 对于例7.1, $R^2 = 0.8542$(见上面输出), 这说明这里的自变量可以大约解释85.4%的因变量的变化. R^2越接近1, 回归就越成功. 由于R^2有随着变量数目增加而增大的缺点, 人们对其进行修改, 因此, 计算机输出还有一个修正的R^2(adjusted R-squared). 对于例7.1, 它等于0.8517. 当然, 它和R^2有类似的意义. 此外, 计算机还计算了一个在零假设下有F分布的检验统计量, 它是用来检验回归拟合好坏的(零假设是因变量和自变量没有关系). 例7.1的F检验的p值也是2×10^{-16}. 这个F检验的H_0为"该回归至少有一个系数(斜率)显著", 这里只有一个自变量, 因此, 前面关于斜率β_1的检验显著就等价于这里F显著, 因此它们的p值相等, 但当回归有至少两个自变量时, 这个F检验的p值就和系数的t检验的p值不同了. 这个F检验也需要上面关于ϵ的各个条件的满足.

7.3.2 多个数量自变量的线性回归

和刚才简单的回归模型类似, 一般的有k个(定量)自变量$x_1, x_2, ..., x_k$的对因变量y的线性回归模型为(称为**多元回归, multiple regression**)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon,$$

这里$\beta_0, \beta_1, ..., \beta_k$称为回归系数. 对计算机来说, 对多个自变量进行回归和一个自变量的情况类似, 只不过多选自变量就是了, 计算机也会自动输出相应的检验结果. 而这些检验也只有在前面说的关于误差项ϵ的各种假定成立时才有意义.

当选定一个模型, 并且用数据来拟合时, 并不一定所有的变量都显著, 或者说并不一定所有的系数都有意义. 软件中一般都有一种一边回归, 一边检验的所谓**逐步回归(stepwise regression)**方法. 该方法或者从只有常数项开始, 逐个地把显著的变量加入(向前逐步回归, forward), 或者从包含所有变量的模型开始, 逐步把不显著的变量减去(向后逐步回归, backward), 也可以为有加有减的双向逐步回归. 这在各种软件都可以实现. 注意不同方向逐步回归的结果也不一定相同. 比方说, 如果一组变量和另一组变量都提供了类似的信息, 这时选择哪一组都有

道理. 还需要注意的是, 逐步回归选择变量的准则也是可以挑选的, 不同的准则会导致不同的结果. 这里R的默认选项为向后逐步回归(backward), 而其默认准则为AIC[1].

例7.1 (数据: bschool.txt, 继续) 现在除了用SalaPostMBA作为因变量之外, 把剩下的三个变量都作为自变量进行回归. 所用的代码为

```
a=lm(SalaPostMBA~.,w);summary(a)
```

该代码和a=lm(SalaPostMBA~SalaPreMBA+Tuition+GMAT,w);summary(a)等同, 当用数据中所有其余变量作为自变量时, 可以用"SalaPostMBA~." 来代替代码中的"SalaPostMBA~SalaPreMBA+Tuition+GMAT". 得到的输出为

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -25.37493   32.91139  -0.771   0.4439
SalaPreMBA    2.38124    0.32345   7.362 8.73e-10
Tuition       0.25759    0.14088   1.828   0.0728
GMAT          0.02820    0.06347   0.444   0.6586
Multiple R-squared: 0.8631,     Adjusted R-squared: 0.8557
F-statistic: 117.6 on 3 and 56 DF,  p-value: < 2.2e-16
```

这里给出了三个系数, 按照t检验, 有一个变量GMAT很不显著. Tuition不是很显著. 可以试着用逐步回归并对残差做Shapiro-Wilk正态性检验:

```
b=step(a);summary(b)
shapiro.test(b$res)
```

得到的回归输出为

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.0657     6.7006  -1.651   0.1041
SalaPreMBA    2.4757     0.2419  10.233 1.61e-14
Tuition       0.2604     0.1397   1.863   0.0676
Multiple R-squared: 0.8626,     Adjusted R-squared: 0.8578
F-statistic: 178.9 on 2 and 57 DF,  p-value: < 2.2e-16
```

这个逐步回归去掉了GMAT, 保留了Tuition. 用y, x_1, x_2代表变量SalaPostMBA, SalaPreMBA, Tuition, 结果的回归方程为

$$y = -11.0657 + 2.4757x_1 + 0.2604x_2.$$

此外, 残差的Shapiro-Wilk正态性检验的p值为0.3126, 因此似乎没有足够证据拒绝正态性假设.

[1]Akaike Information Criterion, 又称为赤池信息准则, 它既要考虑残差平方和要小, 又要考虑待估计参数不能太多(模型的简单化).

除了上面输出之外, 还可以用语句anova(b)输出**方差分析(analysis of variance, anova)表**:

```
Analysis of Variance Table
Response: SalaPostMBA
            Df Sum Sq Mean Sq  F value  Pr(>F)
SalaPreMBA   1  34648   34648 354.3092 < 2e-16
Tuition      1    340     340   3.4718 0.06758
Residuals   57   5574      98
```

方差分析表把因变量与回归无关的总变化平方和(用$\sum_{i=1}^{n}(y_i-\bar{y})^2$表示)分解为若个独立的归因于各个变量及残差的平方和, 在前面提到的ϵ独立同正态分布等条件下, 这些平方和有χ^2分布, 它们(在分别除以各自的自由度之后)的比例有F分布, 这样, 在和残差平方和比之后, 就有了若干F检验, 如果和残差相比显著, 则说明这个变量所解释的变化显著(不能算为随机误差). 从这个表可以看出计算过程. 比如$(34648/1)/(5574/57)=34648/97.789\approx 354.3$, 其中的数字都出现在表中第一和第三行, 这是F统计量的值, 有自由度(1,57), 然后算出p值约等于$0(2\times 10^{-16})$.

7.3.3 自变量中有定性变量的线性回归

例7.2(数据: artif2.txt) 这个数据有三个变量: y, x, u. 其中y和x为数量变量, u为定性变量(有A、B两个水平). 只能够点出y和x的散点图, 图7.4为这样的散点图. 其中左边的是对所有数据, 中间是为$u=A$的部分数据, 右边是为$u=B$的部分数据.

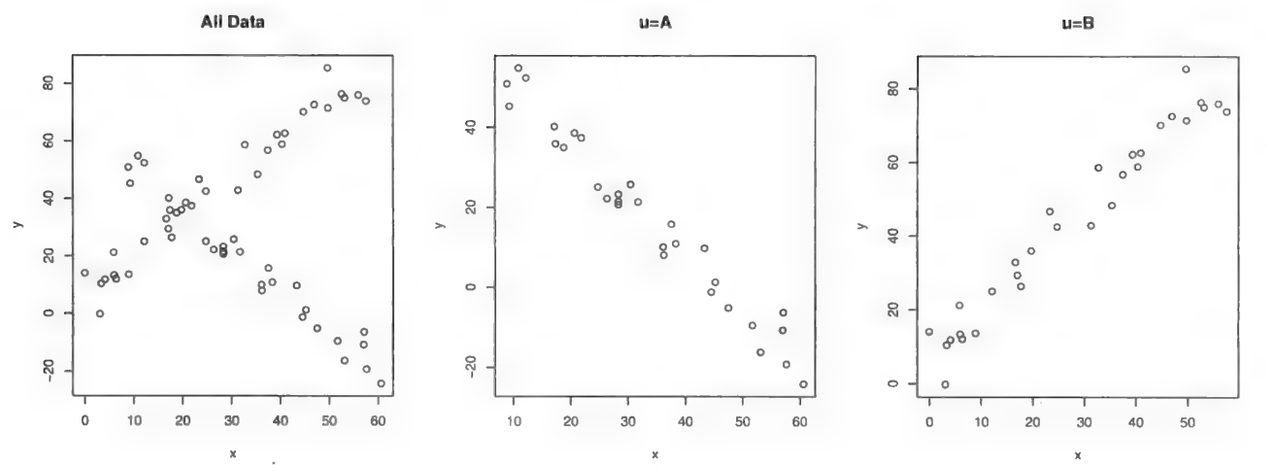

图 7.4 例7.2的x和y的散点图, 左图用了所有数据, 中图只用$u=A$的部分数据, 右图用$u=B$的部分数据.

该图是用下面代码画的(第一行是读数据):

```
w=read.table("artif2.txt",header=TRUE);par(mfrow=c(1,3))
```

```
plot(y~x,w,main="All Data")
plot(y~x,w[w$u=="A",],main="u=A")
plot(y~x,w[w$u=="B",],main="u=B")
```

从图7.4的三个散点图可以看出, 这里实际上是两个不同的回归问题, 看来, 预期的回归直线的截距和斜率都依变量u取值的不同而不同. 我们要选择的模型为

$$y = \mu + (\beta + \beta_i)x + \alpha_i + \epsilon,\ i = 1, 2,$$

这里β是x的固有斜率, α_1和α_2分别为u取A和B时分别对y的效应, 它们是两个数值, 利用一些代数知识, 可以知道, 单独来讲, α_1 或α_2及β_1或β_2是不可估计的, 但它们的差是可以估计的, 因此, 一般需要有约束条件. 一些软件的默认约束条件是设其中一个为零, 比如$\alpha_1 = 0$, 这样在$i = 1$时, 截距为$\mu + \alpha_1 = \mu + 0 = \mu$, 在$i$为其他值时, 截距为$\mu + \alpha_i$. 也有的约束条件是$\sum_i \alpha_i = 0$等等. 无论设定什么约束条件, 一些差, 比如$\alpha_1 - \alpha_2$ $(i \neq j)$或$\beta_1 - \beta_2$ $(i \neq j)$是可以估计的, 不会因为约束条件不同而改变. 而β_i $(i = 1, 2)$为x分别与$u = A$ $(i = 1)$和$u = B$ $(i = 2)$的交互作用, 也就是说, 当变量u取不同值时, 不但截距要变(出现α_i), 而且斜率也要变, 增加β_i. 通过R软件的回归语句(这里模型"y~x*u"等同于"y~x+u+x:u", 而"x:u"代表考虑两个变量的交互作用.):

```
a=lm(y~x*u,w);summary(a)
```

得到下面的输出, 各项检验在正态性假定之下是显著的(p值非常之小, 为10^{-16}的量级),

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  63.43677    2.13594   29.70   <2e-16
x            -1.38886    0.05811  -23.90   <2e-16
uB          -56.07639    2.68658  -20.87   <2e-16
x:uB          2.71269    0.07588   35.75   <2e-16
Multiple R-squared: 0.9692,     Adjusted R-squared: 0.9675
F-statistic: 586.8 on 3 and 56 DF,  p-value: < 2.2e-16
```

在R的这个输出中, 没有显示的uA(即$u = A$的效应α_1)和x:uA(即β_1)被软件设为0, 因而不显示. 对照前面的模型, 这个输出意味着参数的估计为(在原来参数符号上面加"帽子", 注意有些"估计"仅有相对意义)

$$\hat{\mu} = 63.44,\ \hat{\beta} = -1.39,\ \hat{\alpha}_1 = 0,\ \hat{\alpha}_2 = -56.10,\ \hat{\beta}_1 = 0,\ \hat{\beta}_2 = 2.71$$

这产生了分别相应于$u = A$ $(i = 1)$和$u = B$ $(i = 2)$的两条直线;

$$y = 63.44 + (-1.39 + 0)x + 0 = 63.44 - 1.39x\ (i = 1),$$

$$y = 63.44 + (-1.39 + 2.71)x - 56.10 = 7.34 + 1.32x\ \ (i = 2).$$

图7.5显示了在图7.4左面散点图上的这两条回归直线, 是用下面代码画的:

```
plot(y~x,w);abline(a$coe[1],a$coe[2])
abline(a$coe[1]+a$coe[3],a$coe[2]+a$coe[4])
```

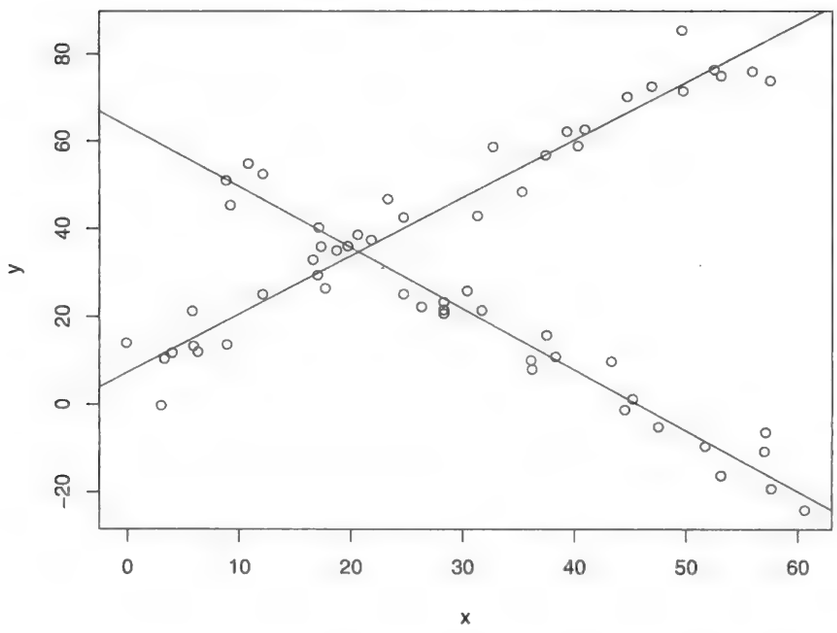

图 7.5　例7.2的x和y的散点图以及两条回归直线.

预测, 训练集, 测试集和交叉验证.

回归的一个重要目的是预测, 即给了一个新的数据, 再套用回归拟合出来的模型算出预测的因变量的值. 在前计算机时代, 要把数据手工代入数据, 进行计算. 现在, 这一切都很简单, 下面就例7.2的数据予以说明. 假定原先的回归结果存在"a"中(回顾前面的代码a=lm(y~x*u,w)). 用代码new=data.frame(u=c("A","B","B"),x=c(47,6,45))创造一个只有u和x的具有3个观测值的名为"new"的数据(注意这里变量的名字x和u要和原先数据的一致, 而次序无关):

```
> new
  u  x
1 A 47
2 B  6
3 B 45
```

只要输入代码predict(a,new), 就可得到三个预测的y值:

```
> predict(a,new)
       1        2        3
-1.83985 15.30332 66.93237
```

一般回归输出中的拟合值(可以用`a$fit`来查看)就是用模型来拟合原来得到该模型的数据的结果(即`predict(a,w)`). 统计模型中的所谓"拟合"好坏, 就是看该模型和用来估计该模型的数据(这种数据叫做**训练数据集, training set**, 简称**训练集**)的适合程度. 比如决定系数R^2就是衡量拟合好坏的一个度量. 模型的建立不仅仅是为了一个数据, 而必须能够解释更广泛的数据. 所谓**过拟合(overfitting)**就是模型和训练集的拟合很好, 但是对其他数据集不合适. 为了客观地评价模型的好坏, 人们往往把一部分数据作为训练集来建立模型, 而另一部分数据作为**测试数据集, (testing set**, 简称**测试集**)来检验模型的误差. 这种方法叫做**交叉验证(cross validation)**. 有时需要进行k折交叉验证(k-fold cross validation), 即把数据分成k份, 每次拿$k-1$份作为训练集, 用剩下的一份作为测试集, 重复k次, 得到k个误差作出平均, 以避免仅用一个测试集可能出现的偏差. 显然, 交叉验证的方法适用于各种模型之间的比较.

思考一下:

1. 例7.2如果没有给出变量u, 就很麻烦了, 这是因为该数据根本不能用一条回归直线描述. 请讨论该例.
2. 在有一个因变量及只有一个k个水平的定性自变量的情况, 得到的是k个截距(y等于k个常数). 如果有若干定量自变量和一个k个水平的定性自变量, 在没有交互作用情况下, 就会产生k条平行回归直线. 如果有若干定量自变量和两个分别有k和q个水平的定性自变量, 在没有交互作用时, 就产生$k \times q$条平行回归直线. 请感兴趣的读者考虑有交互作用的情况.
3. 例7.2实际上体现了不同模型混合在一起的情况. 比如, 一些变量对于不同性质的地区应该服从不同的规律, 即应该用不同的模型来描述, 这时, 表示地区的变量就起了例7.2中的变量u的作用.
4. 如果在平面上有n个点, 我们可以按照横坐标自小到大的次序用$n-1$个折线把它们连接起来, 这时, 该折线就是我们的模型(拟合曲线), 拟合的残差平方和为0, 而R^2及调整的R^2均为1. 这是完美的拟合, 但你会这样做吗?
5. 决定系数R^2的确描述拟合好坏, 但仅此而已.

7.3.4　Logistic 回归

前面回归的因变量为定量变量. 但是如果因变量为取两个值的定性变量, 前面介绍的回归模型就无法解决了. 这实际上是一个分类问题. 在这一节, 通过例7.3来介绍另一种回归, 即logistic回归(logistic regression).

例7.3 脊柱数据(column.2C.dat). 该数据的自变量(V1, ..., V6)为6个生物力学特征, 全部都是关于这些特征的数量变量. 这个数据来源于Frank and

Asuncion (2010) [1]. 我们的研究目的之一是根据这些特征把患者划分两类: 正常(100人, 代码为NO—normal), 不正常(210人, 代码为AB—abnormal). 在这个数据中变量V6的第116个观测值是明显的异常值, 它会影响拟合运算, 因此, 在下面的分析中, 把column.2C.dat的第116个观测值用V6对其他变量回归的方法插补. 代码如下:

```
w2=read.table("column.2C.dat")
ch=lm(V6~.,w2[-116,])
w2[116,6]=predict(ch,w2[116,-6])#50.79539
```

这就将原来的418.54换成了50.79539.

对于一个有两个结果的随机试验, 最简单的概率模型就是Bernoulli试验及Bernoulli分布, 那里假定成功的概率为p, 失败的概率为$1-p$. 二项分布就是由多次Bernoulli试验导出的. 在实际生活中, 有各种不同的其他因素干扰试验结果, 这样成功和失败的概率就不是固定的, 而是其他变量的一个函数. 假定自变量向量为X, 那么一个简单的函数为

$$\ln\left(\frac{p}{1-p}\right) = X^T\beta,$$

式中, β为待估计系数向量. 这和简单回归函数$y = X^T\beta$不同, 方程左边的$\ln(p/(1-p))$不是可观测的变量, 而是假定的背景分布(Bernoulli分布)的一个参数, 因此不能用简单回归的方法来解. 这个回归模型因为其左边函数被称为logit函数而叫做logistic回归模型, 为广义线性模型(generalized linear model, glm)的一个特例. 广义线性模型是关于指数族的一组线性模型, 包括前几节介绍的简单正态线性模型、probit回归模型、Poisson对数线性模型等成员. 在R中可用函数`glm()`来处理. 和logistic回归类似, 结果也往往类似的一个回归模型为probit回归, 其模型也是基于成功概率为p的Bernoulli试验, 其形式为

$$p = \Phi(X^T\beta) \quad \text{或者} \quad \Phi^{-1}(p) = X^T\beta,$$

式中, p为Bernoulli试验的成功概率, Φ为标准正态累积分布函数. 显然, logistic和probit回归都是试图把取值范围为整个实数轴的$X^T\beta$和取值为$[0,1]$区间的p联系起来: 在logistic模型左边的$\ln(p/(1-p))$取值范围和右边$X^T\beta$一样, 都是整个实数轴, 而probit模型左边的p和右边的$\Phi(X^T\beta)$都是取值于$[0,1]$区间的.

下面的代码用logistic回归来拟合column.2C.dat数据(不包括读入数据):

```
#logistic回归在glm中属于binomial族,默认连接函数(link)为logit函数:
a=glm(V7~.,w2,family="binomial")
b=step(a)  #做逐步回归筛选变量
summary(b) #输出回归系数
#由于拟合结果是给每个观测值一个概率值, 下面以0.5作为分类界限:
```

[1]数据可从网站http://archive.ics.uci.edu/ml/datasets/Vertebral+Column下载. 来自Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

```
z=(predict(b,w2,type="response")>0.5)
u=rep("NO",310);u[!z]="AB" #把预测结果转换成原先的值(AB或NO)
(zz=table(w2[,7],u))       #2乘2矩阵,对角线外的数目为分错的数目
(sum(zz)-sum(diag(zz)))/sum(zz)#计算错判率
```

Logistic回归的估计系数及近似正态z检验的结果和分类结果展示在下面两个表中.

逐步回归筛选变量之后的logistic回归结果输出

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-14.7432	3.2037	-4.60	4.19e-06
V1	0.0832	0.0242	3.45	0.000571
V2	-0.1622	0.0359	-4.52	6.10e-06
V3	0.0273	0.0195	1.40	0.162357
V5	0.1049	0.0227	4.62	3.85e-06
V6	-0.1702	0.0234	-7.29	3.15e-13

此外, 在输出中AIC=190.9793. 由上表可知训练出来的模型为

$$\ln\left(\frac{p}{1-p}\right) = -14.7432 + 0.0832\mathrm{V1} - 0.1622\mathrm{V2} + 0.0273\mathrm{V3} + 0.1049\mathrm{V5} - 0.1702\mathrm{V6}.$$

变量V4在逐步回归中被淘汰. 在下表中, 行代表正确的类, 列代表模型判断的类, 对角线外为错判个数. 误判率为0.1483871.

	AB	NO
AB	186	24
NO	22	78

有一个需要注意的是, 由于拟合结果是给每个观测值一个概率值作为拟合值, 我这里按照其大于或不大于0.5作为分类的界限, 这种分法不一定科学, 因为把有病说成没病的损失要大于把没病说成有病的损失. 至于用什么阈值作为分类界限, 必须有一个明确的损失标准. 在实际应用中, 必须注意这一点.

7.3.5 自变量为数量变量时的分类: 经典判别分析

一般的回归是指因变量为定量变量的情况, 而logistic回归主要处理因变量为只取两个值的定性变量情况. 如果作为因变量的定性变量取多于两个值的时候可以用本节介绍的**线性判别分析(linear discriminant analysis)**来建模. 判别分析根据自变量来预测因变量的类型. 判别分析和前一节的logistic回归的目的都是分类, 与其他分类(classification)的目的是一样的. 这里介绍的判别分析开发得比较早, 属于经典的多元分析统计的内容. 它无论在名称上还是在思路上, 均和后面要介绍的算法建模的分类方法有所不同. 这里的判别分析和前面介绍过的回归的思路也有所不同. 注意, 这里的判别分析的自变量只能是定量变量.

判别分析的原理并不复杂, 可以简单地描述如下. 如果作为属性变量的因变量有k个取值(即观测值应该分成k类), 而自变量包含p个变量, 那么每一个观测值就是p维空间的一个点. 整个训练集的各个点(假定有n个点)就按照已知的类别在p维空间中形成共有n个点的k个点群. 那么, 对于一个未知类别的点, 如果离哪一群近, 就可以分到哪一群. 当然还有如何定义"远近"或距离等问题. 下面将要用的R函数所基于的Fisher判别法就是为了使得每一点群内部的点尽可能接近, 而使各群之间尽量分开, 利用了线性代数中的特征值和特征向量的工具, 把原先的p维空间投影到能够把各群最能够分开的低维空间上. 这使得分类更加有道理. Fisher判别法没有像其他线性判别方法那样明确要求假定数据有多元正态总体. 由于篇幅有限, 其他判别方法就不做详细介绍了. 这里仅仅就Fisher线性判别分析一种方法, 通过经典的鸢尾花例子介绍如何通过计算机得到结果及对结果的解释.

例7.4 (数据: iris.txt) 这是鸢尾花(iris)的数据. 该数据给出150个鸢尾花的萼片长(sepal length)、萼片宽(sepal width)、花瓣长(petal length)、花瓣宽(petal width)以及这些花分别属于的种类(Species), 共五个变量. 萼片和花瓣的长宽为四个定量变量, 而作为因变量的种类为分类变量(取三个值: Setosa、Versicolour、Virginica). 鸢尾花为法国的国花, 其萼片也是绚丽多彩的, 和向上的花瓣不同, 花萼是下垂的. 这三种鸢尾花很像, 人们试图建立模型, 根据萼片和花瓣的四个度量来把鸢尾花分类. 这里三种鸢尾花各有50个观测值.

由于鸢尾花数据已经在R软件里面, 省去了输入数据的语句. 运用程序包MASS[1]中的线性判别分析的函数lda()先对所有数据建模:

```
library(MASS)
(a=lda(Species~., iris))
```

得到的结果中有两个线性判别函数的系数(是与Fisher降维方法有关的特征向量中的头两个, 它们把数据从四维空间降到二维):

```
Coefficients of linear discriminants:
                    LD1         LD2
Sepal.Length  0.8293776  0.02410215
Sepal.Width   1.5344731  2.16452123
Petal.Length -2.2012117 -0.93192121
Petal.Width  -2.8104603  2.83918785
```

用下面代码把这二维图点出来(图7.6)

```
z=as.matrix(iris[,1:4])%*%a$scal
plot(z[,1],z[,2],pch=c(rep(19,50),rep(5,50),rep(17,50)),
xlab = "first linear discriminant",
ylab = "second linear discriminant")
```

[1]Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

```
legend("top",pch=c(19,5,17),c("setosa","versicolor","virginica"))
```

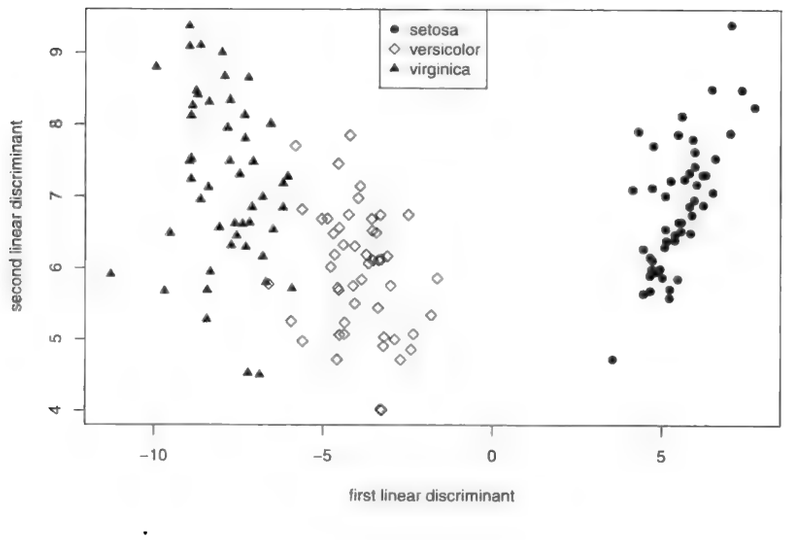

图 7.6　把例7.4对鸢尾花数据分类时把训练集四维空间的点投影到二维空间的结果.

从图7.6可以看出, setosa鸢尾花离另外两种很远, 而versicolor和virginica鸢尾花则比较接近, 容易划分错.

上面的结果是把所有的数据都当成训练集, 没有测试集, 为了进行验证, 我们在每种鸢尾花中随机选择一半(25个)作为训练集(下标用samp表示), 另一半做测试集(下标用-samp表示), 然后用另外25个用来建模. 三种鸢尾花一起, 训练集和测试集各有75个观测. 由于选择测试集和训练集的过程是随机的, 每次计算的结果也不同. 在R软件中, 对于不同的随机数种子, 我们得到的结果也不尽相同. 下面是做这个简单验证(谈不上交叉验证)时的代码:

```
set.seed(1010)
samp=c(sample(1:50,25),sample(51:100,25),sample(101:150,25))
a=lda(Species~., data=iris, subset=samp);
pred=predict(a,iris[-samp,])$class
table(iris[-samp,5],pred)
```

得到测试集分类结果(列为模型预测的, 行为真实的)表

	pred		
	setosa	versicolor	virginica
setosa	25	0	0
versicolor	0	23	2
virginica	0	1	24

有3个错分的, 错分率为$3/75 = 0.04$.

7.4 现代分类和回归: 机器学习方法

前面介绍的回归和分类(判别)模型是可以写成公式的. 但是另外一些回归和分类的方法是体现在算法之中, 其具体形式是计算机程序, 这些方法广泛用于机器学习或数据挖掘之中. 算法模型适用范围比经典的统计模型更加广泛. 由于现在经典模型也要经过计算机软件实现, 因此, 广义地说, 算法模型实际上包含了经典模型, 只不过由于算法模型与经典模型的发展过程及思维方式很不相同, 人们不那么说而已. 算法建模主要发展于最近二十年, 它得益于不断进步的计算机技术. 如果说起源于前计算机时代的经典统计目前大大受惠于计算机的发展, 那么没有计算机, 就不可能产生算法建模.

在处理巨大的数据集上, 在对付被称为维数诅咒的巨大变量数目时, 在无法假定数据的任何分布背景的情况下, 在面对众多竞争模型方面, 算法建模较经典建模有着不可比拟的优越性. 在实际需要拉动下产生和发展的算法建模有着广泛的应用及理论前景.

这里介绍的每个方法都可做回归和分类, 由于它们的产生起因大都是经典统计基本上无能为力的分类问题, 这里也对每种方法先介绍分类, 再介绍回归. 下面分别通过一些数据例子来说明这两方面的方法. 最后还可以对各种方法通过交叉验证进行比较.

例7.5 住房数据(Housing) 该数据可以从网上下载[1], 它有14个变量, 是波士顿郊区506个区域(城镇)的各种统计数据, 说明如下:

变量名	意义	变量名	意义
CRIM	人均犯罪率	DIS	到市中心加权距离
ZN	大面积土地的比例	RAD	到高速路的方便指数
INDUS	非商业面积比例	TAX	每$10000的税率
CHAS	是否接近Charles河(1或0变量)	PTRATIO	学生教师比例
NOX	氮氧化物浓度	B	黑人比例指数
RM	每房平均屋子数目	LSTAT	低阶层人的比例
AGE	1940年前自住房的比例	MEDV	自住房中位数房价

注: ZN为超过25000平方英尺居住土地比例, MEDV的单位为千美元, B的计算公式为$1000(\text{黑人比例} - 0.63)^2$.

其中除了CHAS为哑元(1=接近河,0=否则)之外都是数量变量. 把该数据的中位数房价看成因变量, 其他作为自变量. 因此, 这是一个回归问题. 既可以用经典方法, 也可以用现代机器学习的回归方法处理. 我们将比较各种方法在这个数据上的优劣.

[1]数据网址为http://archive.ics.uci.edu/ml/datasets/Housing. 来自Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

例7.6 皮肤病数据(Dermatology1.txt, Dermatology.txt) 该数据可以从网上下载[1], 该数据有35个变量, 366个观测值, 变量用V1, V2,..., V35表示. 其中前11个为临床属性, 而前10个都是取0,1,2,3的整数(0代表不存在, 1到3代表程度, 越大越显著), 而V11(家族病史)取0或1值; 后面从V12到V33为病理属性, 也是取0,1,2,3的整数(0代表不存在, 1到3代表程度, 越大越显著); V34为年龄; V35为鳞状疾病的类型. 这是个皮肤科数据, 目的是确定Eryhemato鳞状疾病(Eryhemato-Squamous Disease)的类型, V35取值1, 2, 3, 4, 5, 6, 分别代表六种疾病(psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, pityriasis rubra pilaris). 数据文件Dermatology.txt是原始数据, 而Dermatology1.txt是补了年龄(V34)的8个缺失值之后的(下面要用的)数据. 把该数据的疾病类型(V35)看成因变量, 其他作为自变量. 显然, 这是一个分类问题. 由于自变量除了一个二分变量之外, 都是数量变量, 因此传统的线性判别分析也可以使用, 我们最后会比较各种方法在这个数据上的优劣.

例7.7. 蘑菇数据(agaricus-lepiota1.txt, agaricus-lepiota.txt) 该数据可以从网上下载[2], 该数据有23个变量, 8124个观测值, 变量用V1, V2, ..., V23表示. 其中V1为能否食用, 水平"e"(edible)代表可食用, 水平"p"(poisonous)代表有毒; 其余变量都是分类变量, 表示各种蘑菇各部位的形状、颜色、气味、生长特点、生长环境等属性, 全部用字母表示其水平(最多12个水平). 数据文件agaricus-lepiota.txt是原始数据, 而agaricus-lepiota1.txt是补了(V12)的缺失值之后的(下面要用的)数据. 此外, 由于V17只有一个水平, 对建模不起作用. 下面处理时该数据的V1(能否食用)看成因变量, 其他作为自变量. 这是一个因变量只有两个水平的分类问题. 由于自变量全部是分类变量, 经典的判别分析完全不可用, 即使是可以处理少数定性变量的logistic回归在这里也无能为力(三四个定性自变量就无法运行了). 这只能用现代分类方法来处理.

注意, 后面几节对这些数据的分类或者回归结果是对训练集的, 也就是用其自己建立的模型来预测本身, 对模型更科学的判断应该是交叉验证. 通过交叉验证来对分类模型预测效果的评价及各种模型的比较, 我们将在后面集中讨论.

7.4.1 决策树

决策树是本节后面要介绍的其他方法的一个基础. 决策树所能处理的问题非常广泛, 直观易懂, 容易解释, 这是传统统计所不可比拟的. 后面要介绍的boosting和随机森林称为**组合方法**. 几乎所有组合方法的重要研究一开始都是以决策树为基本组件来实现的, 它们大大增进了模型的预测精度.

---

[1]数据网址为http://archive.ics.uci.edu/ml/datasets/Dermatology. 来自Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[2]数据网址为http://archive.ics.uci.edu/ml/datasets/Mushroom. 来自Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

1. 决策树简介, 分类树

决策树的一个突出特点是其再现了人类做决策的过程. 下面用例7.6皮肤病数据(Dermatology1.txt)例子来说明决策树的意义和原理. 图7.7就是根据这个例子所建立的决策树.

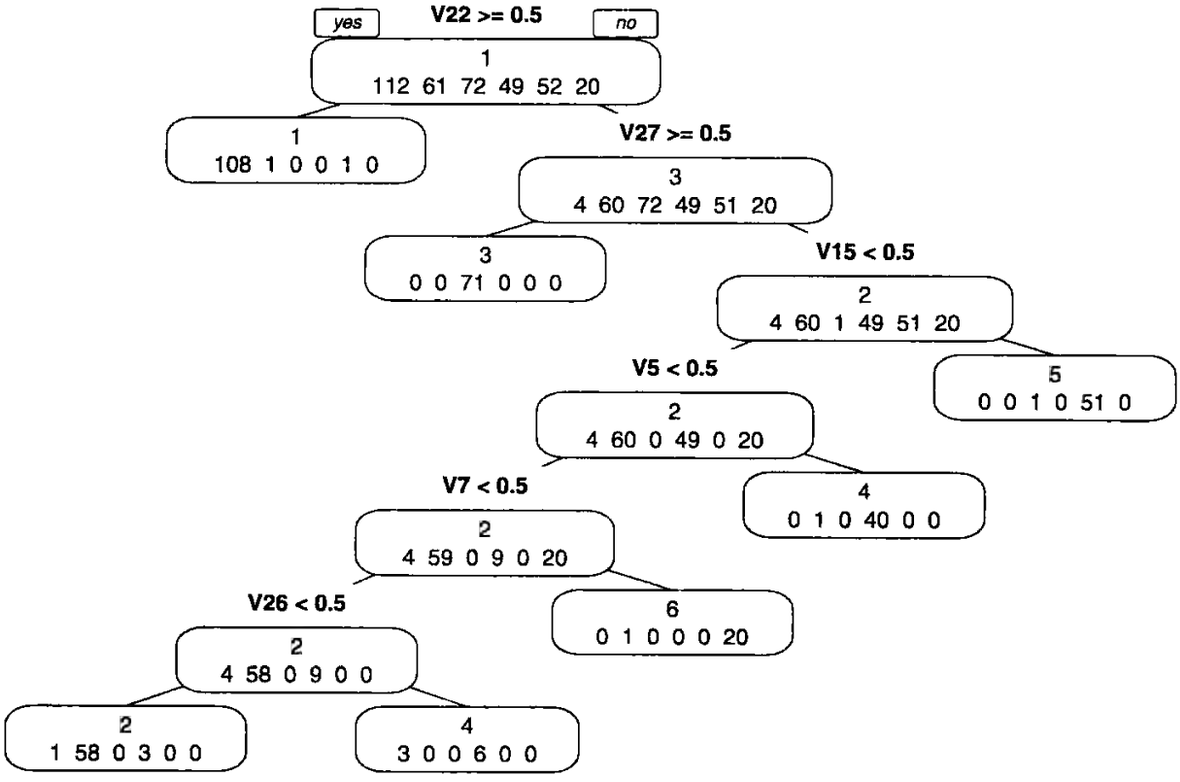

图 7.7　用例7.6数据建立的判别疾病种类(类别代号为1,2,3,4,5,6)的决策树.

该树是用决策树程序包rpart[1]及相配的画图程序包rpart.plot[2]的函数产生的, 具体语句(包括输入数据)如下:

```
w=read.table("Dermatology1.txt",header=T);w[,35]=factor(w[,35])
library(rpart.plot)
(a=rpart(V35~.,w)) #使用全部变量, 用决策树拟合全部数据并打印输出
rpart.plot(a,type=1,extra=1) #画决策树 图7.7
```

除了图中的决策树之外, 细节可以参看输出的打印结果:

[1]Terry M Therneau and Beth Atkinson. R port by Brian Ripley. Note that maintainers are not available to give advice on using a package they did not author. (2012). rpart: Recursive Partitioning. R package version 3.1-53. http://CRAN.R-project.org/package=rpart.

[2]Stephen Milborrow (2012). rpart.plot: Plot rpart models. An enhanced version of plot.rpart.. R package version 1.3-0. http://CRAN.R-project.org/package=rpart.plot.

```
n= 366
node), split, n, loss, yval, (yprob)
      * denotes terminal node
  1) root 366 254 1 (0.31 0.17 0.2 0.13 0.14 0.055)
    2) V22>=0.5 110   2 1 (0.98 0.0091 0 0 0.0091 0) *
    3) V22< 0.5 256 184 3 (0.016 0.23 0.28 0.19 0.2 0.078)
      6) V27>=0.5 71   0 3 (0 0 1 0 0 0) *
      7) V27< 0.5 185 125 2 (0.022 0.32 0.0054 0.26 0.28 0.11)
       14) V15< 0.5 133  73 2 (0.03 0.45 0 0.37 0 0.15)
         28) V5< 0.5 92  33 2 (0.043 0.64 0 0.098 0 0.22)
           56) V7< 0.5 71  13 2 (0.056 0.82 0 0.13 0 0)
            112) V26< 0.5 62   4 2 (0.016 0.94 0 0.048 0 0) *
            113) V26>=0.5 9   3 4 (0.33 0 0 0.67 0 0) *
           57) V7>=0.5 21   1 6 (0 0.048 0 0 0 0.95) *
         29) V5>=0.5 41   1 4 (0 0.024 0 0.98 0 0) *
       15) V15>=0.5 52   1 5 (0 0 0.019 0 0.98 0) *
```

决策树就像一棵从根长出来的树(这里是倒长的, 也有横着长的). 最上面一个叫做**根节点(root node)**, 占据那里的变量为V22(称为**拆分变量**, 后面会解释为什么首先是考虑V22), 在那里的数据按照因变量(V35)的1到6种类别各有:112, 61, 72, 49, 52, 20个(全部数据), 而且标明, 如果不继续, 那么类别1最多(因此在根节点标出"1"). 这时根据V22是否大于等于0.5来做下一步决策, 如果"是"(yes)则走向左边, "不是"(no)则走向右边(一般都按照"yes"往左, "no"往右的习惯); 当走向左边时(V22大于等于0.5的)数据就少了一些, 而且标出1到6种类别各有:108, 1, 0, 0, 1, 0个了, 这时, 由于类别1最多, 因此决策为类别1(有两个分错的), 决策树这个分支就结束了, 而且这个节点就称为**叶节点或终节点(leaf node or terminal node)**, 换言之: 满足V22大于等于0.5的数据最终判断为类别1(在打印输出中, 终节点有"*"号标明); 从根节点往右走(当V22小于0.5), 就进入另一个节点, 那里1到6种类别各有:4, 60, 72, 49, 51, 20个, 最多的是被标明的类别3, 但这种混杂情况很难做决策, 因此这个节点不能为终节点, 称为**中间节点(internal node)**. 那里的拆分变量为V27, 然后根据V27是否大于等于0.5再进分叉. 如此下去, 决策树就长成了. 这个决策树有7个终节点. 其中有6个节点都有些误分的类型. 用语句`table(w[,35],predict(a,w,type="class"))`可以得到下面的表

	1	2	3	4	5	6
1	108	1	0	3	0	0
2	1	58	0	1	0	1
3	0	0	71	0	1	0
4	0	3	0	46	0	0
5	1	0	0	0	51	0
6	0	0	0	0	0	20

其中列号代表真实的类, 行号代表决策树模型所划分的类, 对角线上的数目是正确划分的数目, 而对角线之外的为误分的. 比如第1类观测值被正确划分为第1类的数目为108, 而对第4类观测值被错误划分为第1类的数目为3. 很容易算出误分率为0.03278689, 在总共366个观测值中有12个误分的. 这个决策树仅仅用了34个自变量中间的6个.

决策树的节点上的变量可能是各种形式的(连续、离散、有序、分类变量等等), 一个变量也可以重复出现在不同的节点. 一个节点前面的节点称为**父节点(母节点或父母节点, parent node)**, 而该节点为前面节点的**子节点(女节点或子女节点, child node)**, 并列的节点也叫**兄弟节点(姊妹节点, sibling node)**.

如何挑选拆分变量呢? 以分类为例, 一开始的数据可能包含有若干类, 一般按照下面原则:

步骤1. 如果数据已经只有一类了, 或某一类占绝大部分了(或者按照某停止生长准则), 则该节点为叶节点. 否则进行下一步.

步骤2. 寻找一个变量使得依照该变量的某个条件把数据分成纯度较大的两个(或几个)数据子集. 而用其他变量所划分的子集不如该变量划分得那样纯. 也就是说, 根据某种局部最优性来选择变量. 然后对于其子节点回到步骤1.

上面步骤中说的"纯度"如何定义? 也就是说用什么度量标准来根据数据在某节点选择变量? 不同的软件有不同的标准, 但原理是类似的, 结果不会有多大差别. 另一个问题是让决策树不断地长下去直到无法增长为止, 还是适可而止? 这涉及剪枝问题, 人们总是希望模型既有效又简单. 一般统计软件都有关于剪枝的默认准则, 本书采用的就是默认值(未加改动). 此外, 变量可能是分类变量, 也可能是有序变量或者连续变量. 如果拆分变量是分类变量, 则在其各个水平中找到最优的(使得数据变得最纯) 的水平(或水平组合)作为拆分原则; 如果拆分变量是数量的, 也是寻求一个值, 使得大于或小于该值最能纯化数据. 当然, 我们的目标既可能是分类, 也可能是回归. 例7.6是分类例子, 对于回归例子, 拆分变量的选择则为诸如使得残差平方和最小等准则, 而终节点的决策就是那里余下观测值中因变量的均值.

下面再用决策树来拟合传统统计无法处理的例7.7的蘑菇数据(agaricus-lepiota1.txt)的全部观测值. 所用代码为(包括输入数据):

```
w=read.table("agaricus-lepiota1.txt",header=T)
library(rpart.plot)
(a=rpart(V1~.,w))  #使用全部变量, 用决策树拟合全部数据并打印输出
rpart.plot(a,type=1,extra=1) #画决策树
(z0=table(w[,1],predict(a,w,type="class")))
z0;(E0=(sum(z0)-sum(diag(z0)))/sum(z0))
```

图7.8为得到的决策树.

还输出了打印的决策树细节:

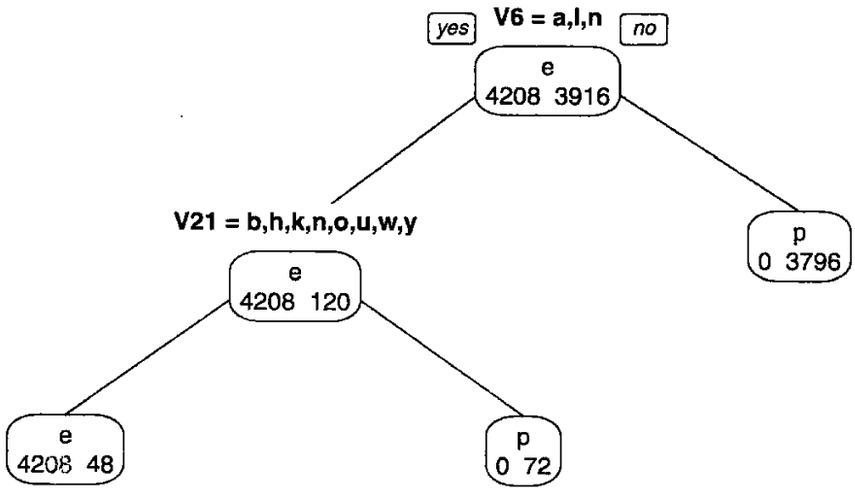

图 7.8 用例7.7数据建立的判别蘑菇是否可食(类别代号为"e","p")的决策树.

```
n= 8124
node), split, n, loss, yval, (yprob)
      * denotes terminal node
1) root 8124 3916 e (0.51797144 0.48202856)
  2) V6=a,l,n 4328  120 e (0.97227357 0.02772643)
    4) V21=b,h,k,n,o,u,w,y 4256   48 e (0.98872180 0.01127820) *
    5) V21=r 72    0 p (0.00000000 1.00000000) *
  3) V6=c,f,m,p,s,y 3796    0 p (0.00000000 1.00000000) *
```

分类结果在下面的2×2矩阵中:

```
     e    p
e 4208    0
p   48 3868
```

在8124个观测中, 一共有48个蘑菇从毒蘑菇错分到可食蘑菇中. 错误率为0.005908419. 后面要介绍的组合方法将对此数据的分析大大改进.

2. **决策树回归: 回归树**

现在用决策树来拟合例7.5的住房数据(Housing)的全部观测值, 使用下面代

码(包括读入数据):

```
w=read.table("housing.txt",header=T)
library(rpart.plot)#同时自动打开rpart
a=rpart(MEDV~.,w);a #计算决策树并输出决策树的细节
rpart.plot(a,type=1,faclen=T) #画出决策树的图
```

得到图7.9及下面关于树细节的输出:

```
n= 506
node), split, n, deviance, yval
      * denotes terminal node
 1) root 506 42716.3000 22.53281
   2) RM< 6.941 430 17317.3200 19.93372
     4) LSTAT>=14.4 175  3373.2510 14.95600
       8) CRIM>=6.99237 74  1085.9050 11.97838 *
       9) CRIM< 6.99237 101  1150.5370 17.13762 *
     5) LSTAT< 14.4 255  6632.2170 23.34980
      10) DIS>=1.5511 248  3658.3930 22.93629
        20) RM< 6.543 193  1589.8140 21.65648 *
        21) RM>=6.543 55   643.1691 27.42727 *
      11) DIS< 1.5511 7  1429.0200 38.00000 *
   3) RM>=6.941 76  6059.4190 37.23816
     6) RM< 7.437 46  1899.6120 32.11304
      12) LSTAT>=9.65 7   432.9971 23.05714 *
      13) LSTAT< 9.65 39   789.5123 33.73846 *
     7) RM>=7.437 30  1098.8500 45.09667 *
```

在这个决策树的每个节点上的数目是该节点处观测值的因变量房价(MEDV)的平均值(单位千美元). 从图7.9可以看出, 回归中对房价最有影响的变量是RM, 还有LSTAT, CRIM和DIS也出现过. 评价模型预测好坏的一个准则为标准化**均方误差(normalized mean squares error, NMSE)**, 定义为

$$\mathrm{NMSE}=\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}=\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2},$$

分子为该模型拟合后的残差平方和($\hat{y}_i$代表对第i观测值的预测), 分母代表用最简单的算术平均$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$(实际上没有用任何模型) 来预测$y$的"残差平方和". 如果NMSE大于或等于1, 说明这个回归模型没有任何意义. 任何有丝毫道理的模型都应该产生NMSE小于1的预测结果. 用下面代码计算NMSE:

```
y0=predict(a,w)
```

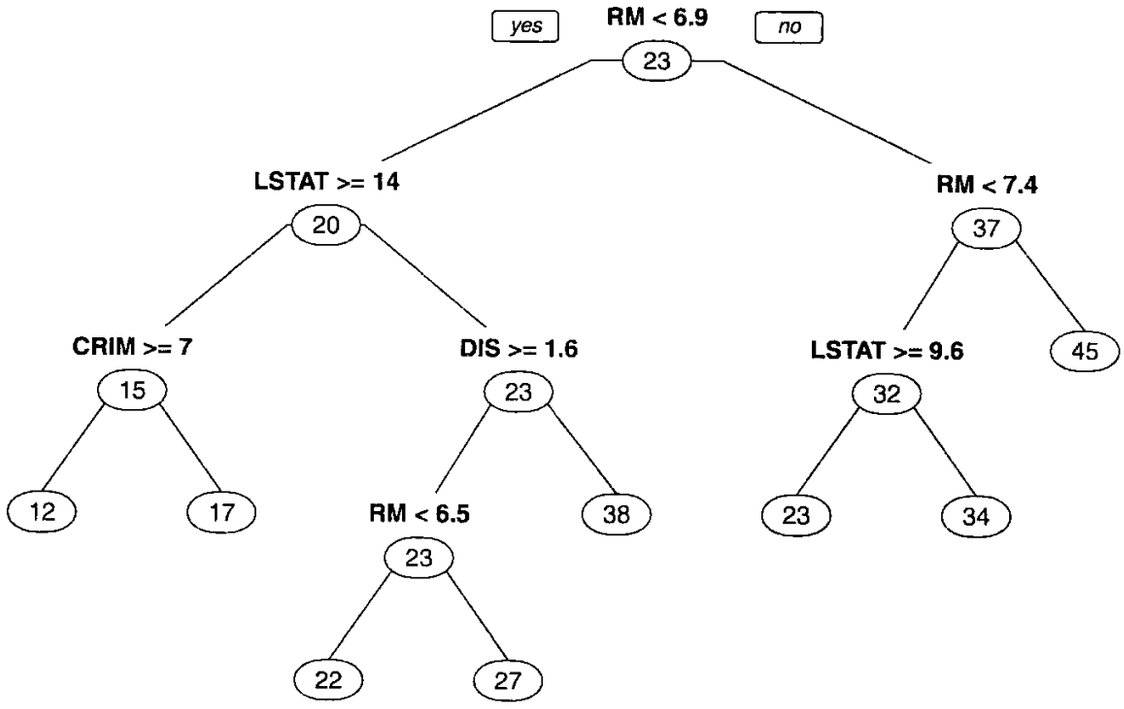

图 7.9　用例7.5数据建立的以中位数房价为因变量的决策树.

```
(NMSE0=mean((w$MEDV-y0)^2)/mean((w$MEDV-mean(w$MEDV))^2))
```

得到NMSE= 0.1924279.

7.4.2　关于组合算法

下面要介绍的几种算法为机器学习组合算法或组合方法, 其目的在于把一些较弱的算法(比如决策树)结合起来产生一个非常精确的预测规则. 为什么组合会得到更好的结果呢? 下面举一个通俗的例子说明.

如果某人欲竞选当地领导, 假定该地有49%的人不支持他. 那么, 每随机问一个人, 都有约49%的可能不选他(我们假定该地选民总数很大, 这样, 每问一个人就近似地相当于一个Bernoulli试验, 相应的概率$p = 0.49$). 如果从该地随机选择1000人来投票, 按照简单多数当选的原则, 那么他不被选上的概率是多少呢? 假定这次投票中不选他的票数服从参数为1000和0.49的二项分布, 容易计算, 这1000人中有超过半数的人(至少501人) 不选他的概率约为0.2532(可用代码`1-pbinom(500,1000,.49)`得到), 远远小于某一个人不选举他的概率0.49. 这类决策例子直观地表现在图7.10中. 该图给出了在个体数目为n时(个体是随机选出的, 在大总体中可近似地看成放回抽样), 个体做某项决策的概率p(横坐标)和个体数目为n的群体按照少数服从多数的投票原则做出该项决策的概率(暂时用p_g表

示, 纵坐标)之间关系的点图(图中的S型曲线). 图中对角线用作对照, 而竖直与水平的两条点状虚线分别标明了这两个概率均为0.5时的位置. 可以看出, 在p小于50%时, 样本量越大, 群体决策概率p_g相对于p越小, 样本量很大时p_g接近于0; 而在p大于50%时, 样本量越大, p_g相对于p越大, 样本量很大时p_g接近于1.

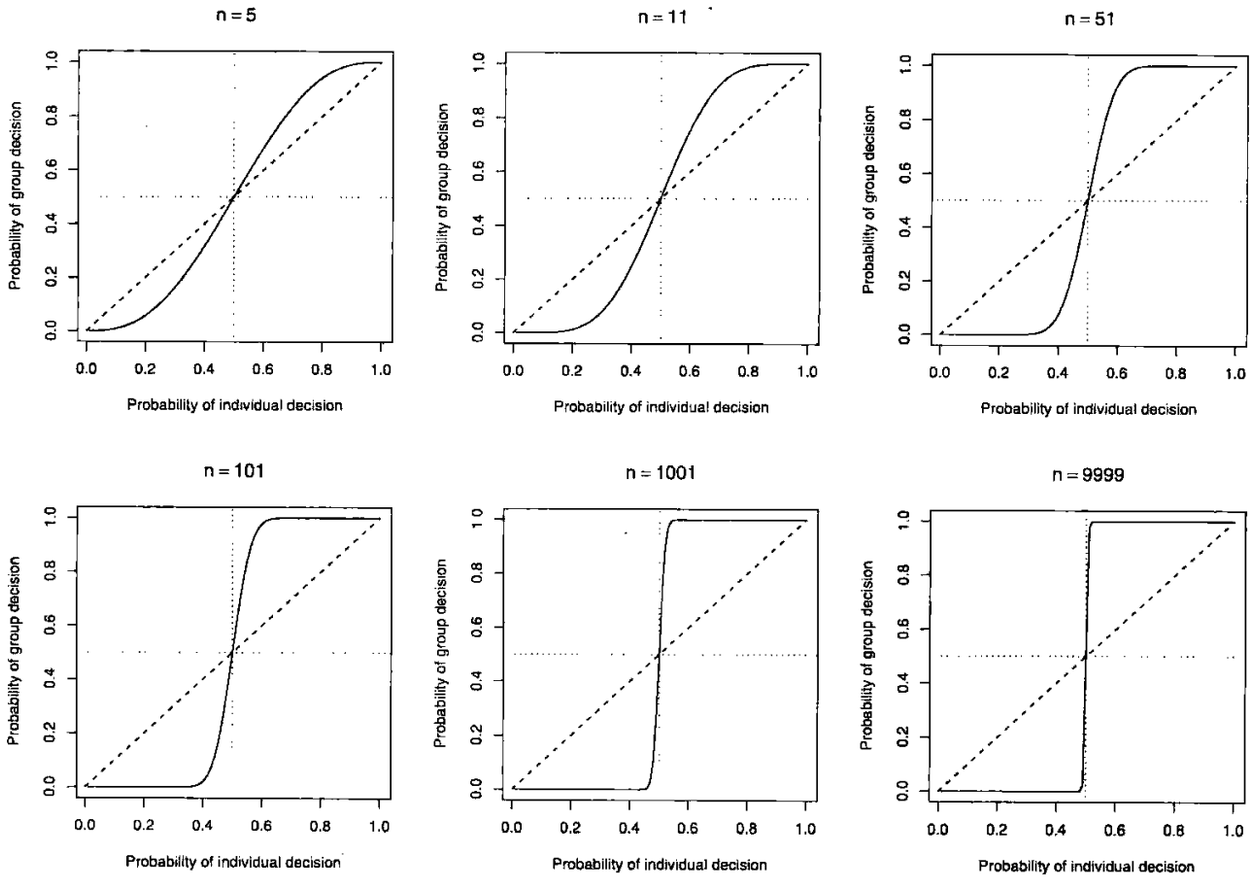

图 7.10 假定个体决策为Bernoulli试验, 个体决策概率(横坐标)与少数服从多数原则下个体数目为n(n分别为5, 11, 51, 101, 1001,9999) 的群体投票决策概率(纵坐标)的点图(S型曲线).

现在考虑分类问题(回归问题类似), 对于组合方法来说, 这时的基本统计模型也称为**分类器(classifier)**. 前面介绍过的分类树和判别分析都是分类器. 假定有许多竞争模型或方法来进行分类, 如果每个方法比随机挑选稍微强一点, 也就是说比用扔硬币要强, 那么每个基本分类器出错误概率应该小于50%. 这时, 类似于刚才所说的选举例子, 用一个分类器的结果, 不如用许多分类器"投票"的结果要可靠. 这种利用多个模型的方法, 对于回归也适用, 只不过不用投票, 而是对不同结果进行某种平均. 这种组合多个分类器或回归模型(称为基本模型)来得到结果的方法就称为**组合方法(ensemble method, ensemble learning, meta algorithm)**. 当然, 图7.9所描述的情况只能是对组合方法出错率的一个简单化类比. 满足Bernoulli试验性质的分类器可能并不存在, 但多个模型的某些形式的组合确实能够大大减少出错率. 下面我们介绍两种把决策树作为基本学习器(基本模型)的组合方法: boosting和随机森林.

7.4.3 Boosting

1. Boosting**简介**, adaboost**分类**

这里介绍的**adaboost**是boosting的一种, 是一种组合方法, 这里用分类树作为基本学习器. Adaboost (adaptive boosting的简写)可以译为自适应助推法, 但我们更愿意用简明的英文缩写adaboost. Adaboost是一种迭代式的组合算法, 目的是分类. 所用的基础分类器(这里用决策树)一开始可能较弱(即出错率较高), 然后, 随着迭代的进行, 不断地通过自助法(bootstrap)加权再抽样, 根据产生新样本来改进分类器, 每一次迭代时都针对前一个分类器对某些观测值的误分缺陷加以修正, 通常的做法是在(放回)抽取样本时对那些误分的观测值增加权重(相当于对正确分类的减少权重), 这样在新的样本中就可能有更多的前一次分错的观测值, 再形成一个新的分类器进入下一轮迭代, 作为结果, 这些观测值在训练模型时就有了更大的代表性, 增加了对这类观测值的正确划分的可能性. 而且在每轮迭代时都对这一轮产生的分类器给出错误率, 最终结果由各个阶段的分类器的按照错误率加权(权重目的是惩罚错误率大的分类器)投票产生. 这就是"自适应". Adaboost的缺点是对奇异点或离群点可能比较敏感, 但其优点是对过拟合不那么敏感. 这里用的程序包是adabag [1], 该程序包包含了adaboost的boosting()(也就是老版本的adaboost.M1()函数, "adaboost.M1"是方法的名称)函数.

下面对例7.6皮肤病数据(Dermatology1.txt)的全部变量和全部观测值用adaboost做分类. 用下面的代码(包括输入数据):

```
w=read.table("Dermatology1.txt",header=T);w[,35]=factor(w[,35])
library(adabag)
set.seed(4410)
a=boosting(V35~.,w)  #旧版本为adaboost.M1(V35~.,w)
z0=table(w[,35],predict(a,w)$class)
z0;(E0=(sum(z0)-sum(diag(z0)))/sum(z0))
barplot(a$importance,cex.name=.7)     #画出变量重要性图
```

这给出了下面输出展示的分类结果, 行是真实类, 列是预测类, 对角线外全部是0, 因此得到误判率为零, 没有一个观测值被错分.

	1	2	3	4	5	6
1	112	0	0	0	0	0
2	0	61	0	0	0	0
3	0	0	72	0	0	0
4	0	0	0	49	0	0
5	0	0	0	0	52	0
6	0	0	0	0	0	20

[1] Alfaro-Cortes, Esteban; Gamez-Martinez, Matias, Garcia-Rubio and Noelia (2011). adabag: Applies AdaBoost.M1, AdaBoost-SAMME and Bagging. R package version 3.0. http://CRAN.R-project.org/package=adabag.

图7.11为变量重要性图.

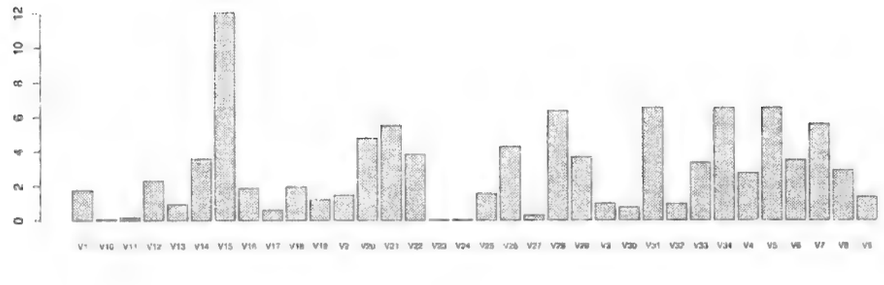

图 7.11　用adaboost拟合例7.6皮肤病数据时的变量重要性图.

从图7.11可以看出, 对于预测最重要的变量为V15, 它代表乳突真皮囊性纤维化(fibrosis of the papillary dermis). 这说明该变量对于识别疾病种类最重要.

下面对例7.7蘑菇数据(agaricus-lepiota1.txt)的全部变量和全部观测值用adaboost做分类. 用下面的代码(包括输入数据):

```
w=read.table("agaricus-lepiota1.txt",header=T)
library(adabag)
set.seed(4410)
a=boosting(V1~.,w)
z0=table(w[,1],predict(a,w)$class)
z0;(E0=(sum(z0)-sum(diag(z0)))/sum(z0))
barplot(a$importance)     #画出变量重要性图
```

这给出了下面输出展示的分类结果, 行是真实类, 列是预测类, 对角线外全部是0, 因此得到误判率为零, 没有一个蘑菇被错分.

```
     e    p
e 4208    0
p    0 3916
```

图7.12为变量重要性图.

从图7.12可以看出, 对于预测最重要的变量为V6, 它代表菌褶依附(gill-attachment)的性状. 这说明该变量对于识别蘑菇是否可食非常重要.

2. Boosting回归

下面用boosting对例7.5住房数据(Housing)的全部观测值和全部变量做回归, 这里用的程序包是mboost[1], 代码为(包括数据输入及求NMSE):

[1]T. Hothorn, P. Buehlmann, T. Kneib, M. Schmid, and B. Hofner (2012). mboost: Model-Based Boosting, R package version 2.1-3, http://CRAN.R-project.org/package=mboost.

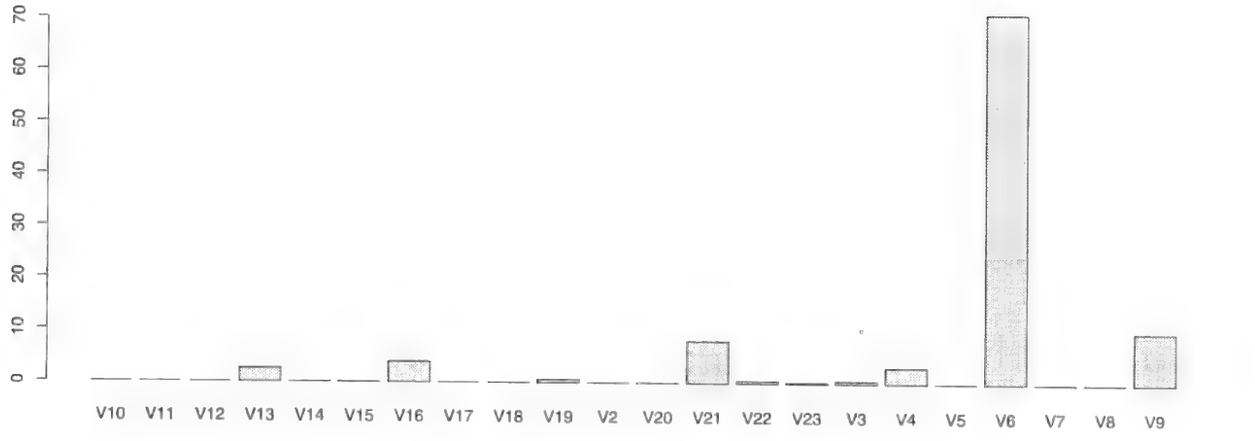

图 7.12　用adaboost拟合例7.7蘑菇数据时的变量重要性图.

```
w=read.table("housing.txt",header=T)
library(mboost)
set.seed(44)
b=blackboost(MEDV ~ .,data =w)
y0=predict(b,w)
(NMSE0=mean((w$MEDV-y0)^2)/mean((w$MEDV-mean(w$MEDV))^2))
```

得到NMSE= 0.0758601.

7.4.4　随机森林

1.　随机森林简介, 随机森林分类

随机森林(random forests)和使用决策树作为基本模型的adaboost有些类似点, 但在每次自助法(boostrap)时都是等权抽样, 这比adaboost简单, 和adaboost不同的是, 在生成树的时候, 每个节点的变量都仅仅在随机选出的少数变量中产生. 因此, 不但每棵树所依据的数据是随机的, 就连每个节点的产生都有很大的随机性. 随机森林让每个树尽量增长, 而且不进行修剪. 随机森林所生成的树的数量大大多于adaboost, 在R中的默认值为500棵树, 而adaboost的默认值为50棵树. 随机森林对于大的数据库很有效率. 它不惧怕很大的维数, 即使是数千变量, 它也不必删除变量, 只要计算机能够承担, 变量多多益善. 随机森林不会过拟合. 它还给出分类中各个变量的重要性. 在一个关于淋巴瘤的基因芯片数据中, 变量个数可以达到4682个, 而样本量仅有81个, 但随机森林可以很好地找到重要的基因(见Breiman, 2001[1]). 这种数据在经典统计中根本无法处理, 因为经典回归分析只能够处理自变量个数大大少于观测值数目的问题, 正如Diaconis &

[1]Breiman, L. (2001) Statistical modeling: the two cultures, *Statistical Science*, Vol. 16, No. 3, 199-231.

Efron(1983) 曾经说过, "统计经验表明, 基于19个变量和仅仅155个数据点来拟合模型是不明智的."[1] 看来这个说法不适用于算法建模.

下面对例7.6皮肤病数据(Dermatology1.txt)的全部变量及全部观测值用程序包randomForest[2]中的随机森林函数做分类. 用下面的代码(包括输入数据):

```
w=read.table("Dermatology1.txt",header=T); w[,35]=factor(w[,35])
library(randomForest)
set.seed(101010)
(a=randomForest(V35 ~ ., data=w,importance=TRUE,proximity=TRUE))
(z0=table(w[,35],predict(a,w)))
#画出变量重要性的8个图
layout(matrix(c(1,2,3,4,5,6,7,7,7,8,8,8),nr=2,byrow=T))
for(i in 1:8)barplot(t(importance(a))[i,],cex.names = 0.5)
```

这给出了下面输出展示的分类结果, 行是真实类, 列是预测类, 没有错分的, 因此得到误判率为零.

```
    1   2   3   4   5   6
1 112   0   0   0   0   0
2   0  61   0   0   0   0
3   0   0  72   0   0   0
4   0   0   0  49   0   0
5   0   0   0   0  52   0
6   0   0   0   0   0  20
```

图7.13为变量重要性图.

图7.13一个有8个图, 上面6个图分别为对因变量疾病类型(V35)的6个类中的每一类判断时的各个变量的重要性(由于图较小, 变量名字没有完全显示, 有兴趣的读者可自行单独做图或者打印数字结果.), 而下面的两个图为各个变量对整个模型的重要性图, 左图是变量对精确度的贡献, 右图是变量作为拆分变量时使数据(在Gini指数的意义上)变纯的贡献.

此外, 随机森林还输出了一行: `OOB estimate of  error rate: 1.91%`, 意思是OOB误差估计为1.91%. 这里OOB是英文"out of bag"的缩写. 由于每次自助法抽样都有一部分观测值没有抽到, 被称为OOB数据集. 显然, 这些观测值就成为天然的交叉验证测试集. 如果按照默认值, 随机森林要建立500棵树, 这样就有500个OOB集作为测试集, 而交叉验证的综合结果就是OOB误差估计. 这样虽然是把整个数据作为训练集来拟合, 但还是进行了大量的交叉验证, 这

[1] Diaconis, P. and Efron, B. (1983). Occam's two razors: the sharp and the blunt. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (R Agrawai and P. Stolorz, eds.) 37-43, AAAI Press, Menlo Park, CA.

[2] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18–22.

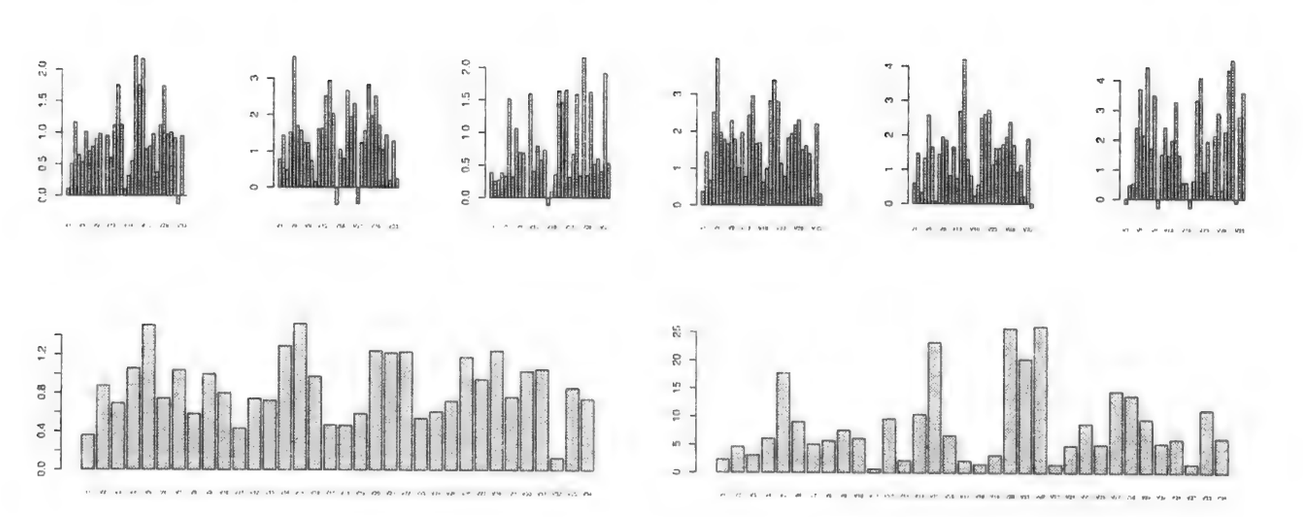

图 7.13　用随机森林拟合例7.6皮肤病数据时的变量重要性图.

里OOB交叉验证的误差为0.0191.

下面对例7.7蘑菇数据(agaricus-lepiota1.txt)的全部变量和全部观测值用随机森林做分类. 用下面的代码(包括输入数据):

```
w=read.table("agaricus-lepiota1.txt",header=T)
library(randomForest)
set.seed(101010)
a=randomForest(V1 ~ ., data=w, importance=TRUE,proximity=TRUE)
z0=table(w[,1],predict(a,w))
z0;(E0=(sum(z0)-sum(diag(z0)))/sum(z0))  #0
#画出变量重要性的4个图
par(mfrow=c(2,2))
for(i in 1:4)barplot(t(importance(a))[i,],cex.names = 0.5)
```

这给出了下面输出展示的分类结果, 行是真实类, 列是预测类, 没有错分的, 因此得到误判率为零, 和adaboost一样.

```
     e    p
e 4208    0
p    0 3916
```

图7.14为变量重要性图.

图7.14一个有4个图, 上面2个图分别为对因变量V1类型(可食还是有毒)的2个类中的每一类判断时的各个变量的重要性, 而下面的两个图为各个变量对整个模型的重要性图, 左图是变量对精确度的贡献, 右图是变量作为拆分变量时使数据(在Gini指数的意义上)变纯的贡献. 这里OOB交叉验证的误差为0.

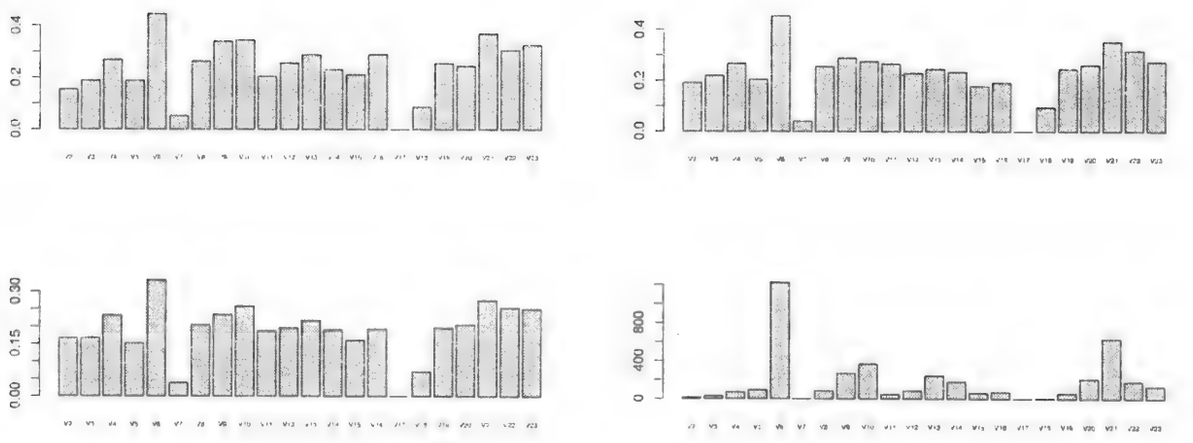

图 7.14　用随机森林拟合例7.7蘑菇数据时的变量重要性图.

2. **随机森林回归**

下面用随机森林对例7.5住房数据(Housing)的全部变量和全部观测值做回归, 代码为(包括数据输入及求NMSE):

```
w=read.table("housing.txt",header=T)
w$CHAS=factor(w$CHAS)
set.seed(1011)
A=randomForest(MEDV~.,data=w,importance=TRUE,proximity=TRUE)
y0=predict(A,w)
(NMSE=mean((w$MEDV-y0)^2)/mean((w$MEDV-mean(w$MEDV))^2))
```

得到NMSE= 0.02310986.

7.4.5　支持向量机

支持向量机(support vector machine, SVM)是一种分类方法, 也可以做回归, 和boosting及随机森林不同, SVM不是基于决策树的组合方法. 它虽然是基于数学模型的, 但充分结合了计算机的算法. 由SVM发展出来的回归方法也称为**支持向量回归(Support Vector Regression, SVR)**. 对于分类问题, SVM是用若干超平面来分割空间以把不同类别的点分开, 在的确可以分开(称为严格线性可分问题)的情况下, 超平面的选择是与其所分割的各类距离最远, 如果允许若干误差(近似线性可分问题), 结果也是一样. 对于线性不可分问题, 可以做变换, 使之成为线性可分问题. 由于线性可分问题通过Lagrange乘子法的解仅仅涉及内积(对偶性质), 线性不可分问题就变成简单地用某个核函数来代替单独变换的内积. 回归用的SVR仅仅是把SVM的思想推广. 该方法之所以称为支持向量机, 是因为确定一个分割超平面的不是所有的点, 而是与超平面最近的若干点, 这些点称为"支持向量"(空间中的点都是向量), 这样就有了支持向量机的名称.

支持向量机主要是为了数量型自变量设计的, 因此对于有大量定性自变量的分类或回归问题不那么适用.

1. 支持向量机分类

下面对例7.6皮肤病数据(Dermatology1.txt)的全部变量和全部观测值用程序包e1071[1]中的svm()函数做分类. 用下面的代码(包括输入数据):

```
w=read.table("Dermatology1.txt",header=T); w[,35]=factor(w[,35])
library(e1071)
a=svm(V35 ~ ., data = w,kernal="sigmoid")
z0=table(w[,35], predict(a,w))
z0;(E0=(sum(z0)-sum(diag(z0)))/sum(z0))
```

这给出了下面输出展示的分类结果, 行是真实类, 列是预测类, 得到误判率为0.01092896.

```
    1   2   3   4   5   6
1 112   0   0   0   0   0
2   0  58   0   3   0   0
3   0   0  72   0   0   0
4   0   1   0  48   0   0
5   0   0   0   0  52   0
6   0   0   0   0   0  20
```

支持向量机对于例7.7的蘑菇数据不适用, 原因是因为蘑菇数据的自变量都是定性变量, 而支持向量机主要是为定量变量设计的.

2. 支持向量机回归

下面用支持向量机对例7.5住房数据(Housing)的全部变量和全部观测值做回归, 这里用的程序包为rminer[2], 代码为(包括数据输入及求NMSE):

```
w=read.table("housing.txt",header=T)
w$CHAS=factor(w$CHAS)
library(rminer)
set.seed(444)
M=fit(MEDV~.,w,model="svm")
y0=predict(M,w)
(NMSE=mean((w$MEDV-y0)^2)/mean((w$MEDV-mean(w$MEDV))^2))
```

得到NMSE= 0.1751441.

---

[1]Evgenia Dimitriadou, Kurt Hornik, Friedrich Leisch, David Meyer and Andreas Weingessel (2011). e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6. http://CRAN.R-project.org/package=e1071.

[2]Paulo Cortez (2011). rminer: Simpler use of data mining methods (e.g. NN and SVM) in classification and regression. R package version 1.1. http://CRAN.R-project.org/package=rminer.

7.4.6　交叉验证比较各个模型

对于一个数据, 可能有很多模型来拟合, 如何衡量和比较模型预测精度呢? 最客观的方法是交叉验证. 交叉验证不需要任何对背景分布等未知的因素做任何的假定. 仅仅是用训练集训练出来的模型来预测没有用来建模的数据(测试集). 这样得出的误差是任何没有学过经典统计的人都能够理解. 交叉验证可以比较任何模型, 无论是经典的还是现代的. 习惯上用5折或10折交叉验证, 这仅仅是习惯. 其实折数越多, 反映出来的误差也越稳定, 但这需要和数据的具体情况相结合. 假定对于一个分类问题做k折交叉验证, 如果因变量有若干个水平, 那么每个水平都需要比较均匀地分成k份, 如果某些水平的观测值少于k个, 那么就可能出现训练集和测试集中各水平不一致的情况. 此外, 对于主要定性自变量的各个水平, 也需要考虑均衡问题. 因此, k的取值需要和具体情况相结合. 此外, 如果要比较若干模型, 这k折的观测值集合应该对所有模型一致.

对于例7.5住房数据(housing.txt)的回归, 我们计算了几种模型10折交叉验证中对训练集预测的标准化均方误差(NMSE), 结果列在下表中:

10折交叉验证训练集的NMSE

模型	NMSE
线性回归	0.2970114
决策树	0.3184500
boosting	0.1496731
随机森林	0.1290078
支持向量机	0.1971914

该结果显示在图7.15中.

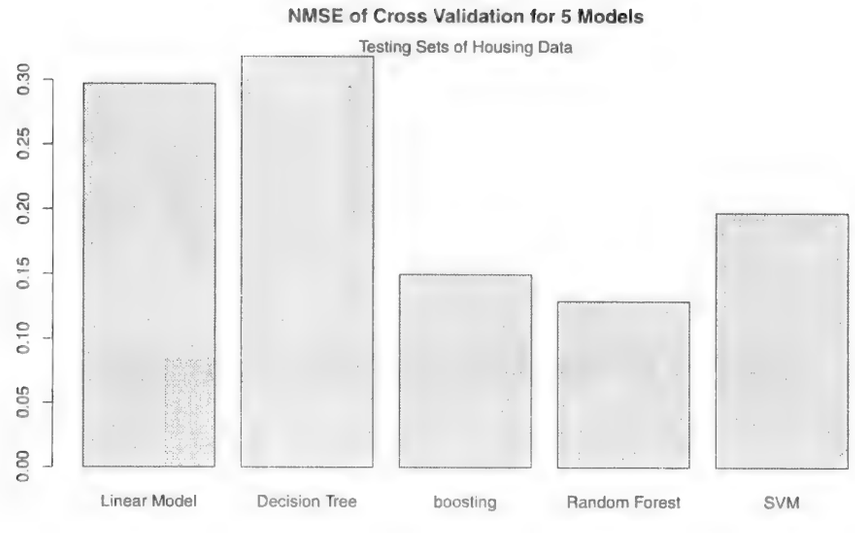

图 7.15　几种模型应用于例7.5时的10折交叉验证中训对练集预测的标准化均方误差(NMSE).

从上面结果可以看出, 随机森林的NMSE最小, boosting次之. 决策树误差最大, 但基于决策树的组合方法以及SVM都优于经典的线性模型, 其中随机森林的NMSE还不到线性模型的一半. 还值得指出的是, 由于线性模型的残差远远不是正态的(Shapiro-Wilk正态性检验的p值为2.2×10^{-16}), 不能做出其满足诸如正态性等假定, 因此, 对于系数的t检验和对于拟合的F检验都失去了理论基础. 只有交叉验证才能显示其价值.

对于例7.6皮肤病数据(Dermatology1.txt)的分类, 我们计算了几种模型10折交叉验证中对训练集预测的误差率(错分类的比例), 结果列在下表中:

10折交叉验证训练集的错分比例

模型	错分比例
线性判别分析	0.03849206
决策树	0.07936508
boosting	0.03809524
随机森林	0.02460317
支持向量机	0.02738095

该结果显示在图7.16中.

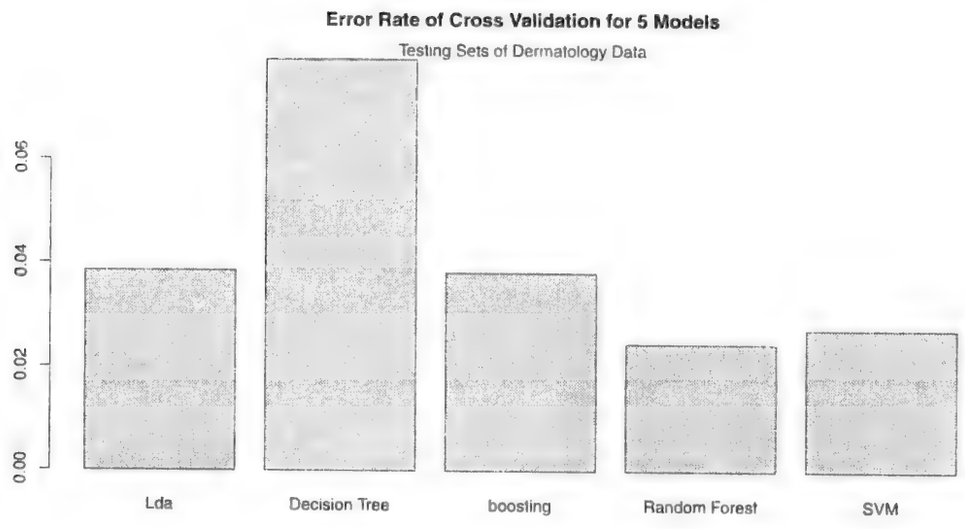

图 7.16　几种模型应用于例7.6时的10折交叉验证中训对练集预测的误差率.

从上面结果可以看出, 随机森林的NMSE最小, SVM次之. 决策树误差最大, 但基于决策树的组合方法以及SVM都优于经典的线性判别分析模型.

对于例7.7蘑菇数据(agaricus-lepiota1.txt)的分类, 我们计算了几种模型10折交叉验证中对训练集预测的误差率(错分类的比例), 结果列在下表中:

10折交叉验证训练集的错分比例

模型	错分比例
logistic回归	失效
线性判别分析	失效
决策树	0.005906803
boosting	0
随机森林	0
支持向量机	失效

从上面结果可以看出, 基于数学模型的logistic回归、线性判别分析、支持向量机都无法应付很多自变量都是定性变量的情况, 而只有决策树及基于决策树的组合方法都能够很好地运作, 其中随机森林及boosting的误差率都是0.

思考一下:
细心的读者可能会想到, 如果用线性回归和线性判别分析作为基础模型来组合, 会不会比用决策树作为基础模型来组合产生更好的结果? 结果并不然, 基于数学模型的这两种方法的组合结果并不理想. 虽然单独来用可能有时强于决策树, 但组合起来改进不大. 读者可以自己思考其原因.

7.5 频数或列联表数据

7.5.1 列联表数据及二维列联表的独立性检验

列联表(contingency table)是一种矩阵形式的表格, 显示的是定性变量不同水平的各种搭配的频数或计数(count). 本节将讨论列联表各个变量之间的关系及对频数变量的建模问题. 下面是一个二维表的例子.

例7.8 眼睛和头发颜色数据(HEColor.txt) 这是关于592个人的头发和眼睛的颜色及他们的性别的数据(Snee, 1974[1]). 头发有四种颜色: 黑色(Black)、金黄(Blond)、棕色(Brown)、红色(Red); 眼睛有四种颜色: 蓝色(Blue)、棕色(Brown)、绿色(Green)、绿棕色(Hazel); 而性别就是女性(Female)和男性(Male); 还有一个变量是每种头发-眼睛-性别组合中的频数(Freq). 下面是这个数据形成的列联表.

Eye	Blue		Brown		Green		Hazel	
Sex	Female	Male	Female	Male	Female	Male	Female	Male
Hair								
Black	9	11	36	32	2	3	5	10
Blond	64	30	4	3	8	8	5	5
Brown	34	50	66	53	14	15	29	25
Red	7	10	16	10	7	7	7	7

[1]Snee, R. D. (1974). Graphical display of two-way contingency tables. *The American Statistician*, 28, 9-12.

这是由下面R代码(包括输入数据)得到的:

```
w=read.table("HEcolor.txt",header=T)
ftable(xtabs(Freq~.,w), row.vars=1,col.var=2:3)
```

二维的列联表又称为交叉表(cross table). 列联表可以有很多维. 维数多的叫做高维列联表. 例7.8中的列联表为$4 \times 4 \times 2$的三维列联表, 列联表的变量一般都是定性变量. 但也可能有一些定量变量与这些定性变量同时存在于原始数据之中, 但数量变量由于取值太多, 不易在表中形成新的维.

研究列联表的一个目的是看这些变量是否相关. 比如例7.8中的头发和眼睛颜色是否相关(不考虑性别时). 下表是把该例的三维表简化成只有头发和眼睛颜色的二维表(用R代码`xtabs(Freq~Hair+Eye,w)`产生),

	Eye			
Hair	Blue	Brown	Green	Hazel
Black	20	68	5	15
Blond	94	7	16	10
Brown	84	119	29	54
Red	17	26	14	14

这时, 检验眼睛颜色相关性的零假设和备选假设为

H_0 : 头发颜色和眼睛颜色这两个变量独立 $\Leftrightarrow$ H_1 : 这两个变量不独立.

这里的检验统计量在零假设下有(大样本时)近似的χ^2分布. 当该统计量很大时或p值很小时, 就可以拒绝零假设, 因而认为这两个变量相关. 对这个检验, 实际上不止有一个χ^2检验统计量. 常用的有Pearson χ^2统计量和似然比(likelihood ratio)χ^2统计量, 它们都有同样自由度的渐近χ^2分布. 这两个统计量的公式将在后面给出, 但不会详细介绍.

就这个例子而言, 根据计算可以得到(对于这两个统计量均有)p值为2.2×10^{-16}, 几乎为0. 计算的R代码为

```
chisq.test(xtabs(Freq~Hair+Eye,w))
```

因此可以说, 头发和眼睛颜色的确相关. 刚才说, 这些χ^2检验是近似的, 那么有没有精确的检验呢? 当然有. 这个检验称为Fisher精确检验, 它所涉及的不是χ^2分布, 而是超几何分布. 计算Fisher统计量得到的p值也很小. 聪明的读者必然会问, 既然有精确检验为什么还要用近似的χ^2检验呢? 这是因为当数目很大时, Fisher检验所基于的超几何分布计算相当缓慢(比近似计算会多很多倍的时间), 而且在计算机速度不快或内存不够时, 根本无法计算. 因此人们多用大样本近似的χ^2检验.

7.5.2　高维列联表和多项分布对数线性模型

例7.8的原始数据是个三维列联表, 这里也可以对三维列联表做各种关于独立性的检验, 也是利用Pearson χ^2统计量和似然比(likelihood ratio)χ^2统计量. 利

用R软件(见下一节), 可以得到下面检验的p值. 这里检验和对两维列联表的检验类似, 但意义有所差异. 这里X, Y, Z分别代表头发颜色、眼睛颜色和性别.

例7.8变量各种独立性的检验

模型记号	零假设	d.f	Pearson检验的p值	似然比检验的p值
(X, Y, Z)	X, Y, Z互相独立	24	0.0000000	0.0000000
(XY, Z)	(X, Y)与Z独立	15	0.1891745	0.1775045
(X, YZ)	X和(Y, Z)独立	21	0.0000000	0.0000000
(XZ, Y)	Y和(X, Z)独立	21	0.0000000	0.0000000
(XZ, XY)	给定X时, Y和Z独立	12	0.4642751	0.4648372
(XY, YZ)	给定Y时, X和Z独立	12	0.1144483	0.1061122
(XZ, YZ)	给定Z时, X和Y独立	18	0.0000000	0.0000000

请读者自己分析上面结果, 至少没有证据拒绝性别(Z)单独和两种颜色的独立性. 上面这些结果都是通过一个所谓的多项分布对数线性模型(multinomial loglinear model)来得到的. 对于列联表(高维或二维)都可以构造(多项分布)对数线性模型来进行分析. 利用对数线性模型的好处是不仅可以更好解释数据, 而且可以增加定量变量作为模型的一部分. 该模型之所以被冠以"多项分布", 是因为把落入列联表各个格子的频数看成是符合多项分布的(参见4.3.1节多项分布).

现在简单直观地通过例7.8的三维表介绍一下对数线性模型. 用m_{ijk}代表三维列联表第i个头发颜色, 第j个眼睛颜色及第k种性别的期望频数(这里i和j取$1, 2, 3, 4$个值之一, k取1或2). 假定列联表格子中的期望频数(各种组合计数)属于多项分布, 该期望频数可以用下面的公式来描述:

$$\ln(m_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k.$$

这就是所谓的(多项分布)对数线性模型. 这里式子右边为头发颜色的第i个水平α_i眼睛颜色的第j个水平β_j性别的第k个水平γ_k对$\ln(m_{ijk})$的综合影响. 这三个影响称为**主效应(main effect)**. 除了主效应之外, 还有可能有**交互效应或交互作用(interaction)**, 交互效应意味着多个变量同时作用时, 对$\ln(m_{ijk})$的效果不是这些变量主效应的简单相加, 而是或者多一些, 或者少一些. 对于例7.8来说, 较完全的模型应该有两个变量的交互效应和三个变量在一起的交互效应, 这种把所有变量的所有效应都列出的模型称为**饱和模型(saturated model)**, 下面就是例7.8的饱和模型:

$$\ln(m_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{jk} + (\beta\alpha)_{ki} + (\alpha\beta\gamma)_{ijk}$$

尽管在模型中可以写上很多效应, 但不一定都有意义. 前面给出的变量各种独立性的检验表中所列的实际上就是对这个饱和模型的各种形式的检验. 细节这里就不介绍了. 上面模型中的关于效应的参数很多, 但它们只有相对意义. 在公式中还有一个截距项μ, 它在这个模型中没有什么意义, 但在软件输出中可能会出现, 这有数学上的原因, 主要是因为各个效应不能单独估计出来, 也就是说, 总有截距混在一起, 说不清截距有多少属于某个效应. 但无论如何, 效应之间的差, 比如

$$\alpha_1 - \alpha_2,\ \alpha_2 - \alpha_3,\ \alpha_3 - \alpha_1,\ \alpha_1 + \alpha_2 - 2\alpha_3$$

等等是完全可以估计的, 因为在这些差中, 截距被减掉了. 在软件中也可以输出那些单独或交互效应的"估计"值, 这里的估计之所以打引号是因为一个效应(以用α表示的变量为例)各个水平的影响是相对的, 因此, 只有事先固定一个参数值(比如设定某$\alpha_i = 0$), 或者设定类似于$\sum_i \alpha_i = 0$这样的约束, 才可能估计出各个α_i的值. 没有约束, 这些参数是算不出来的.

用(多项分布)对数线性模型可以在总观测数固定时估计m_{ijk}的值, 但一般来说多项分布对数线性模型并不是用于预测, 而是通过这个模型做上面提到的各种检验. 下面用只有头发颜色和眼睛颜色两个主效应的$\ln(m_{ij}) = \mu + \alpha_i + \beta_j$模型来拟合例7.8数据, 用R代码

```
library(MASS); a=loglm(Freq~Hair+Eye,w);a$para
```

得到这些参数μ, α_i, β_j的估计:

```
$`(Intercept)`
[1] 2.648531
$Hair
      Black       Blond       Brown         Red
-0.17911627 -0.01706041  0.79474431 -0.59856762
$Eye
      Blue       Brown       Green       Hazel
 0.5067010  0.5296905 -0.7050540 -0.3313375
```

显然, 这些参数的约束条件是$\sum_i \alpha_i = 0, \sum_i \beta_j = 0$. 从输出中可以看出$\mu, \alpha_i$和$\beta_j$的估计为(估计值用"戴了帽子"的$\hat{\alpha}_i$和$\hat{\beta}_j$表示):

$$\hat{\mu} = 2.648531,\ \hat{\alpha}_1 = -0.17911627,\ \hat{\alpha}_2 = -0.01706041,$$
$$\hat{\alpha}_3 = 0.79474431,\ \hat{\alpha}_4 = -0.59856762,$$
$$\hat{\beta}_1 = 0.5067010,\ \hat{\beta}_2 = 0.5296905,\ \hat{\beta}_3 = -0.7050540,\ \hat{\beta}_4 = -0.3313375$$

思考一下:

1. 用例7.8数据拟合的$\ln(m_{ij}) = \mu + \alpha_i + \beta_j$模型实际上是$4 \times 4 = 16$个式子, 每个右边为一个常数. 请想一下模型$\ln(m_{ij}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$一共有多少式子呢?

2. 请大家对多项分布线性模型的注意力集中在有关变量是否独立这个问题上面, 模型本身的预测功能并没有吸引多少注意力. 但是, 下一节的Poisson对数线性模型则正相反, 那里主要应注意模型本身.

3. 在回归中也有交互作用与饱和模型, 请讨论.

7.5.3　Poisson对数线性模型

有的时候, 类似的高维表并不一定满足多项分布对数线性模型. 比如波浪造

成的船舶损害、某地由于AID病死亡的人数、某种病的发病数目、某区域自杀的数目、宇宙飞船被空间粒子击中的次数、某地区的交通事故数等等, 这些都有可能近似地符合Poisson分布的(参见第四章). 下面看一个例子.

例7.9 机器事故(acc2.txt) 这是关于某类机器发生事故次数(Incidents)、机龄(Time, 定量变量)、机器型号(Machine, 两种机器: 1, 2)、操作人(Person, 两类操作工人: 1, 2)的数据.

这个问题显然和列联表的问题不一样, 也不能用前面解决列联表的方法来解决. 可以考虑Poisson对数线性模型. 假定发生事故服从Poisson分布, 但是由于条件不同, Poisson分布的参数λ也应该随着条件的变化而改变. 这里的条件就是所给出的机龄、型号与工人类别等三个变量. 当然, 这里所关心的是这些变量如何影响Poisson分布, 以及这些影响是否显著. 这个模型可以写成

$$\ln(\lambda) = \mu_i + \beta_j + \gamma x,$$

这里λ为常数项, α_i为机器类别($i = 1, 2$分别代表两个水平), β_j为操作者类别($j = 1, 2$代表两个水平), x为连续变量机龄, 而γ为年龄前面的系数. 这里之所以对Poisson分布的参数$\lambda(> 0)$取对数, 是为了使模型左边的取值范围为整个实数.

用例7.9数据拟合Poisson对数线性模型, 这次用R软件来计算, 代码为

```
m=read.table("acc2.txt",header=T)
m$Machine=factor(m$Machine); m$Person=factor(m$Person)
a=glm(Incidents~Time+Machine+Person,family="poisson",data=m);
summary(a)
```

得到下面输出:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.655345   0.385525  -1.700 0.089154
Time         0.005937   0.001662   3.571 0.000355
Machine2     0.416216   0.176388   2.360 0.018291
Person2      0.143591   0.176933   0.812 0.417047
```

这意味着对μ的估计为-0.655345, 对α_i的两个值的"估计"为0.000000, 0.416216, 对β_j的两个值的"估计"为0.000000, 0.143591, 对斜率γ的估计为0.005937.

和前面多项分布对数线性模型类似, 这里的对主效应α_i和β_j的估计只有相对意义, 它们是在一个参数设定为0的约束条件下得到的. 从模型看上去, 对于事故来说, 操作者并不那么重要. 机龄和型号都还显著, 但机龄似乎更重要.

注意, 并不是所有的类似数据用Poisson对数线性模型都适用. 必须大体上有Poisson分布的背景. 一般来说, 在某些固定的条件下, 人们认为某些事件出现的次数服从Poisson分布, 比如在某一个时间段内某种疾病的发生数、显微镜下的微生物数、血球数、门诊病人数、投保数、商店的顾客数、公共汽车到达数、电

话接通数等等. 然而, 条件是不断变化的. 因此, 所涉及的Poisson分布的参数也随着变化. 这也就使得了人们考虑Poisson对数线性模型.

由于Poisson分布只有一个参数λ, 它既是均值又是方差, 但在实际数据中的均值与方差可能不同, 这时, 如果强行用Poisson模型拟合就会产生实际方差大于均值的所谓**过离散(overdispersion)**现象或者实际方差小于均值的**欠离散(underdispersion)**现象. 这时可能需要用有两个参数的模型, 比如负二项分布或gamma分布线性模型等其他模型来拟合. 这有些超出了本书范围, 但过离散和欠离散现象绝非少见.

对数线性模型还有一个问题就是数据中计数为0的数目大大多于其他的整数, 这种数据称为**零膨胀计数数据(zero-inflated count data)**. 这也需要专门对付, 例如, R软件中的程序包`pscl`就有应付零膨胀计数数据的程序.

思考一下:

1. Poisson对数线性模型也会有交互作用, 应该注意到模型变量间的交互作用往往被一些分析者所忽视, 而在许多情况下, 交互作用可能比主效应更显著.
2. 请讨论一下Poisson对数线性模型和多项分布对数线性模型的区别.

7.6 小结

7.6.1 本章的概括和公式

1. 相关

本章介绍了线性相关分析及衡量相关的三个度量: Pearson相关系数、Kendall τ相关系数和Spearman ρ秩相关系数. 其中Pearson相关系数的原理是把每一对观测值(x_i, y_i)中的y_i值到均值$\bar{y}$的距离$y-\bar{y}$与相应的x_i值到$\bar{x}$的距离$x_i-\bar{x}$相乘, 得到$(x_i-\bar{x})(y_i-\bar{y})$. 如果这个乘积为正, 那么说明相对于各自的均值, x_i和y_i的变化趋势一样, 如果这个乘积为负, 那么说明它们的变化趋势相反. 把样本中所有这些乘积相加, 得到和$\sum_i(x_i-\bar{x})(y_i-\bar{y})$. 如果样本中的乘积$(x_i-\bar{x})(y_i-\bar{y})$多数为正, 那么乘积和$\sum_i(x_i-\bar{x})(y_i-\bar{y})$也为很大的正数; 如果这些乘积多数为负, 那么乘积和为绝对值很大的负数; 如果这些乘积的正负号差不多, 那么乘积和就接近于0. 再把它标准化, 就成为取值在-1和1之间的一个量了, 即Pearson相关系数, 其公式为

$$r=\frac{\sum_{i=1}^n(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^n(x_i-\bar{x})^2\sum_{i=1}^n(y_i-\bar{y})^2}}$$

这里$x_1,...,x_n$和$y_1,...,y_n$为两个样本的观测值.

所有这三个相关系数是-1到1之间的数, 越接近1(或-1)就越正(负)相关, 越接近0, 就越不相关. 和这三个相关系数有关的是三个检验统计量(零假设为相关系数为0), 其中与Kendall τ相关系数和Spearman ρ秩相关系数相关的检验属于非

参数检验范畴(与总体分布无关). 注意, 如果拒绝零假设, 即得到相关系数不为0的结论, 但这不意味相关, 因为相关系数很小, 也是不为零. 计算两个变量x和y的样本相关系数(三种选项)的R代码为

```
cov(x, y, method = c("pearson", "kendall", "spearman"))
```

这里默认值为后面method的第一个选项(Pearson相关系数), 如果要得到其他相关系数, 比如Kendall τ, 则可用代码`cor(x,y,method="kendall")`.

对两个变量定义了相关系数之后, 对一组变量可以定义相关阵. 假定该组有p个变量, 它们的相关阵为一个$p \times p$矩阵, 其第ij个元素为第i个变量和第j个变量的相关系数. 对角线元素是变量和自己的相关系数, 等于1. 样本相关阵很容易从R软件得到, 假定w是一个数量变量的数据矩阵(行为观测值, 列为变量), 代码`cor(w)`就给出了样本相关系数阵.

2. 经典线性回归和分类

(a) 经典线性回归分析

对于自变量和因变量(假定有k个自变量和一个因变量)都是定量变量时, 回归模型为

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon,$$

这里$\beta_0, \beta_1, ..., \beta_k$称为回归系数, ϵ为误差项. 本书介绍的是用最小二乘法来得到直线的参数. 在回归时, 如果有误差项为独立同正态分布的假定, 就可以对各个回归系数(t检验)和整个模型的拟合(F检验)进行检验. 当然模型也可能会有两个或两个以上变量的交互作用(比如还有x_i及x_j的交互作用, 上面方程就增加一项$\beta_{ij} x_i x_j$). 此外还有描述拟合的统计量R^2(决定系数), 定义为

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{\sum_{i=1}^{n}(y_i - \bar{y})},$$

它越接近1, 代表拟合越好. 而调整的R^2定义为

$$\bar{R} = R^2 - (1 - R^2)\frac{k}{n - k - 1},$$

这里n为观测值数目, k为自变量的数目. 注意, R^2仅仅描述拟合, 其对于模型其他性质没有多大发言权.

当自变量有定性变量时, 经典回归模型会有所不同. 比如例7.2的含有2个水平的变量u作为自变量的模型就成为

$$y = \mu + (\beta + \beta_i)x + \alpha_i + \epsilon, \quad i = 1, 2$$

这里的α_1, α_2 (β_1, β_2也一样) 只有相对大小可以估计, 因此要设立对它们的约束条件, 比如设一个等于0, 或者它们的和为1等等. 在没有交互作用时, 分析结果时可看它们的相对大小, 比如$\alpha_1 - \alpha_3$和$\alpha_2 - \alpha_3$等.

(b) 两个属性的分类: logistic 回归

当因变量为二水平定性变量时, 把它看成成功概率为p的Bernoulli试验的结果. 但成功概率p为自变量的函数, 于是模型就变成

$$\ln\left(\frac{p}{1-p}\right) = X^T\beta,$$

这里方程右边的解释和线性回归类似.

(c) 自变量为数量变量的分类: 线性判别分析

判别分析利用了有若干变量值的一些已知其所属类别的观测点(训练样本), 并用它们来导出基于这些变量的对未知观测点的分类方法. 有p个变量的一个观测值为p维空间的一个点. 有点就可以定义距离. 判别分析的基本原理是一个点应该属于离它最近的那一类. 为了更好地区分各类, Fisher判别分析在分析距离前先进行投影, 使得各类的投影尽可能分开, 而各类内部尽可能密切. 由于并不是所有变量在进行判别分析时都是重要的, 因此, 可以一边判别, 一边对变量进行筛选, 这就是逐步判别.

判别分析方法很多, 所涉及的公式很烦杂. 不同的距离定义、不同的方法都涉及很多的公式. 我们觉得不引进这些公式既不会妨碍对判别分析概念的理解, 也不会影响对实例的判别. 有兴趣刨根问底的读者请参阅有关的多元分析的出版物. 这里仅就Fisher判别法中如何寻找投影方向的数学予以描述. 记点x在以a为法方向的投影为a^Tx. 而各组数据的投影为(假定有k类, 而第i类有n_i个点)

$$G_i:\ a^Tx_1^{(i)}\cdots a^Tx_{n_i}^{(i)},\ i=1,...,k.$$

将G_m类中数据投影的均值记为$a^T\bar{x}^{(m)}$, 则有

$$a^T\bar{x}^{(m)} = \frac{1}{n_m}\sum_{i=1}^{n_m} a^T\bar{x}_i^{(m)},\ m=1,...,k.$$

记k类数据投影的总均值为$a^T\bar{x}$, 则有

$$a^T\bar{x} = \frac{1}{n}\sum_{m=1}^{k}\sum_{i=1}^{n_m} a^T\bar{x}_i^{(m)}.$$

类间离差平方和为

$$\begin{aligned} SSG &= \sum_{m=1}^{k} n_m(a^T\bar{x}^{(m)} - a^T\bar{x})^2 \\ &= a^T\left[\sum_{m=1}^{k} n_m(\bar{x}^{(m)}-\bar{x})(\bar{x}^{(m)}-\bar{x})^T\right]a = a^TBa, \end{aligned}$$

这里$B=\sum_{m=1}^{k} n_m(\bar{x}^{(m)}-\bar{x})(\bar{x}^{(m)}-\bar{x})^T$. 类内离差平方和为

$$SSE = \sum_{m=1}^{k}\sum_{i=1}^{n_m}(a^Tx_i^{(m)} - a^T\bar{x}^{(m)})^2$$

$$= a^T \left[\sum_{m=1}^{k} \sum_{i=1}^{n_m} (x_i^{(m)} - \bar{x}^{(m)})(x_i^{(m)} - \bar{x}^{(m)})^T \right] a$$
$$= a^T E a$$

这里$E = \sum_{m=1}^{k} \sum_{i=1}^{n_m} (x_i^{(m)} - \bar{x}^{(m)})(x_i^{(m)} - \bar{x}^{(m)})^T$. 根据Fisher方法的原则, 希望寻找方向a使得类间离差平方和SSG尽可能大而类内离差平方和SSE尽可能小, 也就是说, 使得各类内部点尽可能地接近, 而各类之间尽可能分开, 一个途径就是希望

$$\Delta(a) = \frac{a^T B a}{a^T E a} \quad .$$

尽可能大. 而使得$\Delta(a) = a^T Ba/a^T Ea$尽可能大的向量$a$为方程$|B - \lambda E| = 0$的最大特征根$\lambda_1$所相应的特征向量(这是所谓的广义特征值问题), 而最大值就是该特征根λ_1. 记方程$|B - \lambda E| = 0$的全部特征根为$\lambda_1 \geq \cdots \geq \lambda_r > 0$, 相应的特征向量为$v_1, \cdots, v_r$. 而$\lambda_i$的大小可以用来评估判别函数$y_i(x) = v_i^T x (= a^T x)$的效果. 如果记$p_i$为判别能力(效率), 即前面说的贡献率, 有$p_i = \lambda_i / \sum_{h=1}^{r} \lambda_h$, 而$m$个判别函数的累积判别能力或累积贡献率定义为$\sum_{i=1}^{m} p_i = \sum_{i=1}^{m} \lambda_i / \sum_{h=1}^{r} \lambda_h$. 然后, 根据累积贡献率的大小来选择取几组方向. 比如选取两个方向, 得到两个典则判别函数F_1和F_2. 投影之后, 每个观测值根据这两个函数就得到两个坐标, 成为这两个方向所组成的平面中的一个二维点(如图7.6). 然后根据距离各类重心远近来决定任意一个点应该划归哪一类.

3. 现代分类和回归: 机器学习方法

前面的经典回归和分类方法的实施需要许多无法验证的数学假定, 而且都是线性的, 对于自变量中定性变量的适应性很差. 这里介绍的几种现代分类和回归方法没有任何总体分布的限制, 所有的问题都可以是非线性的, 除了支持向量机之外, 这些方法对于处理大量定性变量非常方便. 判别预测方法的好坏及比较不同的方法的预测效果可以用交叉验证来实行. 交叉验证可以用于各种模型. 由于交叉验证不需要任何大学本科概率论与统计的知识, 任何领域的人都能够理解其结果.

由于这里介绍的方法主要是算法, 除了支持向量机之外没有多少数学内容. 这里只做简单概括.

(a) 决策树: 组合方法的基石.

决策树一开始就以处理定性变量及分类问题为出发点的, 而这些是传统统计的弱项. 决策树的要点是选择最能够使因变量观测值变纯或残差平方和最小的拆分变量. 决策树的结果很容易理解, 也很好应用. 虽然单独的决策树的结果可能不如其他一些方法, 但其组合起来则成为可以非常精确预测的组合方法.

(b) 组合方法: boosting和随机森林

我们这里仅介绍了两种组合方法. 它们都基于对样本进行放回再抽样来建立许多决策树. 在抽样上, boosting每次都根据前一棵树的结果调整对观测值的抽样权数, 以使得结果不好的观测值有更大的代表性, 而随机森林则一直都是等权抽

样; 在选择每棵树的拆分变量上, boosting是让所有变量平等竞争, 而随机森林则仅仅在随机选出的若干变量之间竞争, 以增加某些变量的代表性; 在最后的投票上, 随机森林为等权投票, 而boosting是按照各个树的误差大小来加权投票; 在树的数目上, 随机森林要大大多于boosting. 它们都不会过拟合, 为很优秀的分类和回归方法.

(c) 支持向量机的原理

支持向量机虽然对于大量分类变量的自变量不那么有效, 但的确是一个很好的方法. 其数学原理简要概括如下:

1. SVM分类. 假定其目的是把空间中的两类点($y=-1$或$y=1$)用超平面$\boldsymbol{w}^T\boldsymbol{x}+b=0$分开(在所谓严格线性可分的情况下, 存在这样的超平面), 而且希望这个超平面距离两类点的距离最大, 也就是说, 使得隔离带宽$\rho=2/\|\boldsymbol{w}\|$最大. 这等价于用Lagrange乘子法求下式的极小值.

$$L(\boldsymbol{w},b,\alpha)=\frac{1}{2}\|\boldsymbol{w}\|^2-\sum_{i=1}^{n}\alpha_i[y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)-1].$$

根据得到的解$\boldsymbol{w}^*,b^*,\alpha^*$得到最优分割超平面方程$\boldsymbol{w}^{*T}\boldsymbol{x}+b^*=0$. 任意点$(\boldsymbol{x})$的函数值$\boldsymbol{w}^{*T}\boldsymbol{x}+b^*$的符号确定了该点的分类, 或者说判别函数为$\mathrm{sgn}(\boldsymbol{w}^{*T}\boldsymbol{x}+b^*)$. 上面是严格线性可分的情况. 如果允许一些错误, 则称为近似线性可分问题, 结果与此有同样的形式. 可以注意到, 这些结果有下面依赖于内积的对偶(duel)性质: 首先, 这里的建模过程仅仅依赖于训练集点对的内积. 其次, 判别过程仅仅依赖于未知点和训练集中支持向量的内积. 这种依赖于内积的独特的性质使得我们能够解决线性不可分的问题. 对于线性不可分问题, 可能需要一些变换$x\mapsto\Phi(x)$, 这些变换是很难猜到的, 但基于对偶性质, 可以猜测较灵活的核函数$K(\boldsymbol{x}^T,\boldsymbol{x}_j)=\Phi(\boldsymbol{x}_i)^T\Phi(\boldsymbol{x}_j)$而非$\Phi(\cdot)$本身.

2. SVM回归或SVR. 在回归问题中, y不仅仅是-1和1. 令$f(\boldsymbol{x})=\boldsymbol{w}^T\boldsymbol{x}+b$, 希望$y$与$f(\boldsymbol{x})$的离差越小越好, 问题还是归结于求使得$\|\boldsymbol{w}\|^2/2=\boldsymbol{w}^T\boldsymbol{w}/2$最小的$\boldsymbol{w}$, 但约束条件是$\|y_i-f(\boldsymbol{x}_i)\|\le\epsilon$, 这里$\epsilon$为某目标值. 类似于SVM分类, 允许一些误差, 这样就可以把上面的约束放宽为(对于大于0的ξ_i,ξ_i^*) $y_i-f(\boldsymbol{x}_i)\le\epsilon+\xi_i$及$f(\boldsymbol{x}_i)-y_i\le\epsilon+\xi_i^*$, 即得到Lagrange函数

$$\begin{aligned}L(\boldsymbol{w},b,\xi,\alpha,\eta)=&\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_i(\xi_i+\xi_i^*)-\sum_i(\eta_i\xi_i+\eta_i^*\xi_i^*)\\&-\sum_i\alpha_i(\epsilon+\xi_i-y_i+\boldsymbol{w}^T\boldsymbol{x}_i+b)\\&-\sum_i\alpha_i^*(\epsilon+\xi_i^*+y_i-\boldsymbol{w}^T\boldsymbol{x}_i-b).\end{aligned}$$

需要在约束条件$\alpha,\eta>0$下, 解$\min_{\boldsymbol{w},b,\xi}\{\max_{\alpha,\eta}L(\boldsymbol{w},b,\xi,\alpha,\eta)\}$问题. 对于非线性问题, 和前面SVM分类一样, 利用核函数来解决.

4. 频数和列联表数据

本章首先介绍了什么是列联表, 它是若干定性变量的各种可能取值(水平)组合的出现频数表. 研究列联表的主要目的是看这些变量是否相关. 而分析的主要手段是使用χ^2检验. 检验统计量是Pearson χ^2统计量和似然比χ^2统计量. 在零假设(不相关)下, 这两个χ^2统计量都有渐近的χ^2分布. 通过计算机软件, 可以得到所需要的p值. 这里χ^2检验的零假设是二维表中行变量和列变量不相关(检验相关性), 或者是对数线性模型没有意义(检验拟合优度). 实际上二维列联表的相关性检验等价于二维表相应的对数线性模型的拟合优度检验(test of goodness of fit), 即检验模型拟合的好坏.

Pearson χ^2统计量和似然比χ^2统计量是怎么定义的呢? 假定列联表有n个格子, 在例7.8数据中$n = 4 \times 4 \times 2 = 32$个格子. 而各个格子里面的数目(频数)假定为$O_1, ..., O_n$. 根据零假设, 可以算出每个格子数目的期望值$E_1, ..., E_n$. 这里用字母$O$表示观测的值(observed value), 字母E表示零假设下期望的值(expected value). 这样Pearson χ^2统计量Q和似然比χ^2统计量T就分别定义为

$$Q = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \text{ 和 } T = 2\sum_{i=1}^{n} O_i \ln \frac{O_i}{E_i}.$$

直观上说, 如果零假设正确, 则通过零假设得到的期望的E_i不会和观测的O_i差太远. 那么Q和T就不会很大, p值也不会很小, 则不能拒绝零假设. 但如果零假设不对, 那么E_i就会和观测的O_i差很远, 于是Q和T就会很大, 这样就得到很小的尾概率p值, 以至于拒绝零假设.

对于列联表还可以用(多项分布)对数线性模型来描述. 以二维列联表为例, 用m_{ij}表示第(ij)个格子的期望频数, 那么只有主效应的对数线性模型为

$$\ln(m_{ij}) = \alpha_i + \beta_j.$$

这相应于只有主效应α_i和β_j, 而这两个变量的效应是简单可加的. 但是有时, 两个变量在一起时会产生附加的交互效应, 这时, 相应的对数线性模型就是

$$\ln(m_{ij}) = \alpha_i + \beta_j + (\alpha\beta)_{ij}.$$

由于前面对这个模型已有解释, 这里就不重复了.

对于表格中数目有符合Poisson变量的特定意义时(比如例7.9的事故数), 就要考虑是否用Poisson对数线性模型. 如例7.9那样有两个定性变量及一个定量变量的Poisson对数线性模型可以表示为

$$\ln(\lambda) = \mu + \alpha_i + \beta_j + \gamma x.$$

这里μ为常数项, α_i和β_j为两个定性变量的主效应, x为连续变量, 而γ为其系数. 这里之所以对Poisson分布的正参数λ取对数, 是为了使模型左边的取值范围为整个实数轴. Poisson对数线性模型有可能有过离散, 欠离散的问题, 还可能有零膨胀计数问题, 这可能会使得该模型完全失效, 必须注意.

7.6.2　R语句的说明

由于除了交叉验证之外所有关于回归和分类的例子的R代码都在正文中解释了, 这里仅就k折交叉验证如何建立训练集和测试集的做一简要说明. 此外, 对一些列联表数据检验等未在正文明示的R软件使用也在此做一汇总.

1.　把n个下标随机地分为k份

这个比较简单, 下面是个小函数:

```
CV1=function(n=100,k=10,seed=888){ #输入样本量n, 折数k和随机种子seed
z=rep(1:k,ceiling(n/k))[1:n]
set.seed(seed);z=sample(z,n)
mm=list();for (i in 1:k) mm[[i]]=(1:n)[z==i]
return(mm)}  #最后得到的mm的每一个分量mm[[i]]是第i折的下标集
```

以例7.5住房数据(Housing.txt)为例, 下面是包括输入数据在内的对线性回归预测的10折交叉验证程序, 对于其他方法也类似编程. 这里利用了上面的函数.

```
w=read.table("housing.txt",header=T)
w$CHAS=factor(w$CHAS)
(n=nrow(w));k=10;mm=CV1(n,k)
NMSE=rep(0,k)          #建立一些向量以存结果
for(i in 1:k){         #对每一组训练集和测试集做一次, 共k次
m=mm[[i]]              #m为测试集下标集合
a=lm(MEDV~.,data=w[-m,])    #简单线性回归, 这里[-m]为训练集下标集合
y1=predict(a,w[m,])    #对测试集预测
#测试集的NMSE:
NMSE[i]=mean((w$MEDV[m]-y1)^2)/mean((w$MEDV[m]-mean(w$MEDV[m]))^2)}
(MNMSE=mean(NMSE))     #下面输出训练集及测试集的平均NMSE:
```

2.　把n个下标按照定性因变量的类型均衡地随机分为k份

有些因变量的水平(类)很不平衡, 为了使得在交叉验证的每一折中, 每个水平(类)都有相应的代表, 必须把每一类都分成k份. 这个必须具体例子具体分析. 就例7.6皮肤病数据(Dermatology1.txt)为例, 因变量(第35个变量V35)有六个水平, 相应于各水平的观测值数目不那么均匀. 为此, 以10折交叉验证为例, 写下以下程序来产生10个下标集(包括输入数据):

```
w=read.table("Dermatology1.txt",header=T) #输入数据
w[,35]=factor(w[,35])
n=nrow(w);T=length(table(w[,35]));Z=10
#上面n为样本量,T为因变量类别数,Z为折数
d=1:n;dd=list(); e=names(table(w$V35))
```

```
for(i in 1:T)dd[[i]]=d[w$V35==e[i]] #每个dd[[i]]是i类下标集
#下面每个kk[i]是i类中每折的数目:
kk=NULL;for(i in 1:T)kk=c(kk,round(length(dd[[i]])/Z))
set.seed(111);yy=list(NULL,NULL,NULL,NULL,NULL,NULL)
for (i in 1:T){xx=list();uu=dd[[i]];
 for (j in 1:9) { xx[[j]]=sample(uu,kk[i])
 uu=setdiff(uu,xx[[j]])};xx[[10]]=uu
for(k in 1:10)yy[[i]][[k]]=xx[[k]]}
mm=list(NULL,NULL,NULL,NULL,NULL,NULL,NULL,NULL,NULL,NULL)
for(i in 1:Z)for(j in 1:T)mm[[i]]=c(mm[[i]],yy[[j]][[i]])
#mm[[i]]是第i折的测试集下标集合
```

利用了上面输出的下标集合利用线性判别分析(对于其他模型也类似)对皮肤病数据的V35做预测的10折交叉验证的代码:

```
library(MASS);E1=rep(0,Z);for(i in 1:Z){m=mm[[i]];
n1=length(m);a=lda(V35~.,w[-m,])
E1[i]=sum(w[m,35]!=predict(a,w[m,])$class)/n1};mean(E1)
```

3. 和列联表有关的R软件程序

(a) 二维表(考虑例7.8的数据HEcolor.txt)

```
w=read.table("HEcolor.txt",header=T) #输入数据
w1=xtabs(Freq~Hair+Eye,w) #然后建立头发颜色和眼睛颜色的二维表:
chisq.test(w1)    #对两变量相关做卡方检验
fisher.test(w1)   #做Fisher精确检验(对这个问题不推荐因为数据量太大)
```

(b) 高维列联表的各种独立性检验(考虑例7.8数据, 这里X, Y, Z代表头发颜色, 眼睛颜色和性别三个变量)

读入数据:`w=read.table("HEColor.txt",header=T);x=xtabs(Freq~.,w)`

模型记号	零假设	R语句
hline (X, Y, Z)	X, Y, Z互相独立	a=loglin(x,list(1,2,3))
(XY, Z)	(X, Y)与Z独立	a=loglin(x,list(1:2,3))
(X, YZ)	X和(Y, Z)独立	a=loglin(x,list(1,2:3))
(XZ, Y)	Y和(X, Z)独立	a=loglin(x,list(2,c(1,3)))
(XZ, XY)	给定X时, Y和Z独立	a=loglin(x,list(1:2,c(1,3)))
(XY, YZ)	给定Y时, X和Z独立	a=loglin(x,list(1:2,2:3))
(XZ, YZ)	给定Z时, X和Y独立	a=loglin(x,list(c(1,3),2:3))

相应的量(如p值)可以由下面的语句得到

- 自由度(d.f): `a$df`
- 似然比检验统计量T的值: `a$lrt`

- 似然比检验的p值: `pchisq(a$lrt,a$df,low=F)`
- Pearson检验统计量T的值: `a$pear`
- Pearson检验的p值: `pchisq(a$pear,a$df,low=F)`

(c) (多项分布)对数线性模型(考虑例7.8的数据HEColor.txt)

```
w=read.table("HEcolor.txt",header=T) #输入数据
w1=xtabs(Freq~.,w) #然后建三维表:
library(MASS); a=loglm(Freq~Hair*Eye+Sex,w)#主效应和一个混合效应
summary(a);anova(a);a$para #系数
```

输入数据:`w=read.table("HEColor.txt",header=T);x=xtabs(Freq~.,w)` 用下面语句拟合(多项分布)对数线性模型(只有主效应):

(d) Poisson对数线性模型(考虑数据acc2.txt)

```
m=read.table("acc2.txt",header=T) #输入数据
m$Machine=factor(m$Machine); m$Person=factor(m$Person)
a=glm(Incidents~Time+Machine+Person,family="poisson",data=m);
summary(a)
```

7.7 习题

1. 利用例7.2数据(artif2.txt), 把y作为因变量, 仅把x作为自变量进行回归, 画出x, y散点图及回归直线, 结果如何, 请讨论.

2. 利用例7.2数据(artif2.txt), 把y作为因变量, 仅把u作为自变量进行回归, 画出x, y散点图及回归直线, 结果如何, 请讨论.

3. 利用例7.2数据(artif2.txt), 把y作为因变量, 把u和x作为自变量, 不考虑交互作用进行回归, 画出x, y散点图及回归直线, 结果如何, 请讨论.

4. 数据logi.txt是200个不同年龄和性别的人对某项服务产品的认可的数据. 这里年龄是连续变量, 性别是有男和女(分别用1和0表示)两个水平的定性变量, 而变量观点则为包含认可(用1表示)和不认可(用0表示)两个水平的定性变量. 人们想要知道的是究竟年龄和性别对观点有没有影响, 有什么样的影响, 请用本章介绍过的模型表示出这个关系.

5. (数据logi.txt)把性别作为因变量, 把年龄和观点作为自变量, 进行logistic回归, 解释结果.

6. 利用例7.5, 7.6, 7.7数据, 做各种方法的交叉验证(重复产生7.4.6节结果的各种运算).

7. (数据:diabetes.scale.txt)目标: 按照第一列的变量(因变量), 利用其他变量(自变量)分类, 做交叉验证.

8. (数据:svmguide2.txt)目标: 按照第一列的变量(因变量), 利用其他变量(自变量)分类, 做交叉验证.

9. (数据:bodyfat.txt)目标: 按照第一列的变量(因变量), 利用其他变量(自变量)回归, 做交叉验证.

10. (数据:fourclass.txt)目标: 按照第一列的变量(因变量), 利用其他变量(自变量)分类, 做交叉验证.

11. (数据:mpg.txt)目标: 按照第一列的变量(因变量), 利用其他变量(自变量)回归, 做交叉验证.

12. (数据: glass.scale.txt)目标: 按照第一列的变量(因变量), 利用其他变量(自变量)分类, 做交叉验证.

13. 想出你自己设计的一个二维列联表, 总频数不要太大, 用Fisher精确检验和χ^2检验得到结论. 你的零假设是什么? p值是多少? 可否拒绝?

14. 解释关于列联表的多项分布对数线性模型和Poisson对数线性模型在本质上有什么区别.

15. 数据中有一个acc.txt, 它类似于acc2.txt, 只不过没有变量Person(其他的也不尽相同). 用Poisson线性模型来拟合它.

第八章　多元分析

传统的多元分析包括主成分分析、因子分析、聚类分析、典型相关分析、判别分析等内容, 这里变量大都要求为有多元正态分布. 后来, 一些教科书又包括了不那么"经典"的对应分析. 除了前面已经介绍过的线性判别分析之外, 这一章将介绍所有的这些方法.

8.1　寻找多个变量的代表: 主成分分析和因子分析

假定你是一个公司的财务经理, 掌握了公司的所有主要数据, 比如固定资产、流动资金、每一笔借贷的数额和期限、各种税费、工资支出、原料消耗、产值、利润、折旧、职工人数、职工的分工和教育程度等等. 如果让你向有关方面介绍公司状况, 你能够把这些指标和数字都原封不动地摆出去吗? 当然不能. 你必须要把各个方面进行高度概括, 用一两个指标简单明了地把情况说清楚. 其实, 每个人都会遇到有很多变量的数据. 比如全国或各个地区的带有许多经济和社会变量的数据, 各个学校的研究、教学及各类学生人数及科研经费等各种变量的数据等等. 这些数据的共同特点是变量很多, 在如此多的变量之中, 有很多是相关的. 人们希望能够找出它们的少数"代表"来对它们进行描述. 注意, 如果各个变量都独立, 主成分分析和因子分析是没有意义的.

本节就介绍两种把变量维数降低以便于描述、理解和分析的方法: **主成分分析(principal component analysis)**和**因子分析(factor analysis)**. 实际上主成分分析可以说是因子分析的一个特例. 这两种方法的目的是一样的, 都是寻找众多相关变量的少数代表, 这些代表变量, 又称为成分或因子, 都是原先变量的线性组合. 由于代表变量的数目显著地小于原先变量的数目, 数据的维数也就因而降低了.

主成分分析数学较简单, 发展也较早, 因子分析需要的数学假定较多, 理论稍微有些复杂, 但结果可能会比主成分分析更理想. 这一节的目地是找出这些由线性组合而形成的成分或因子, 并且试图解释它们的意义.

8.1.1　主成分分析

为了直观地描述主成分分析降维的过程, 先假定原先数据只是两个变量的观测值, 即二维数据. 如果这两个变量分别由横轴和纵轴所代表, 每个观测值都有相应于这两个坐标轴的两个坐标值, 也就是这个二维坐标系中的一个点. 如果这些数据点形成一个有椭圆形轮廓的点阵, 如二维正态变量的情况[1]. 那么这个椭圆有一个长轴和一个短轴, 称为主轴. 主轴之间是互相垂直的. 在短轴方向上, 数据变化较小, 在长轴方向上, 数据变化较大. 如果两个坐标轴和椭圆的长短轴平行, 那么代表长轴的变量就描述了数据的主要变化, 而代表短轴的变量就描述了数据的

[1]一般地说, 只有在变量近似地服从多维正态分布时, 主成分分析和因子分析的效果才会好, 那时, 多维数据点阵形成多维空间的椭球形状.

次要变化.

但是, 坐标轴通常并不和椭圆的长短轴平行. 因此, 需要寻找椭圆的长短轴, 并进行变换, 使得新变量和椭圆的长短轴平行. 如果长轴变量代表了数据包含的大部分信息, 就用该变量代替原先的两个变量(舍去次要的短轴变量), 降维就完成了. 在极端的情况, 短轴如果退化成一点, 那只用长轴变量就能够完全解释这些点的变化, 这样, 由二维到一维的降维就自然完成了. 图8.1是一个这样的椭圆的示意图. 椭圆的长短轴相差得越大, 降维也越有道理.

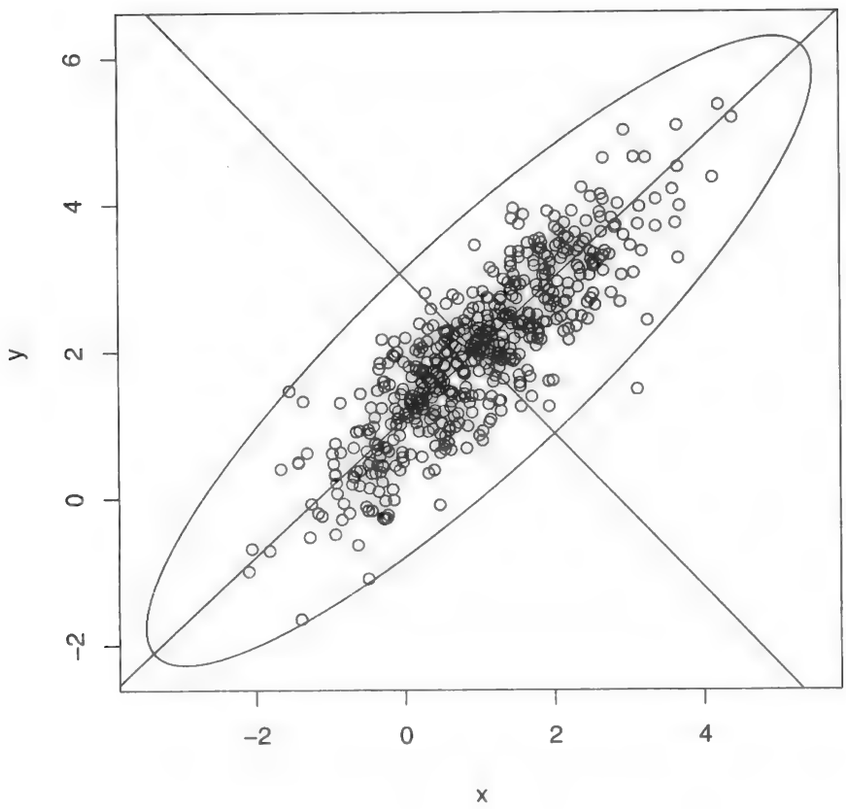

图 8.1　二维空间主成分示意图.

多维变量的情况和二维类似, 也有高维的椭球, 只不过无法直观地看见罢了. 首先把高维椭球的各个主轴找出来, 再用代表大多数数据信息的最长的几个轴作为新变量, 这样, **主成分分析(principal component analysis)**就基本完成了. 注意, 和二维情况类似, 高维椭球的主轴也是互相垂直的. 这些互相正交的新变量是原先变量的线性组合, 叫做**主成分(principal component)**.

正如二维椭圆有两个主轴, 三维椭球有三个主轴一样, 有几个变量, 就有几个主成分. 当然, 选择越少的主成分, 降维就越好. 什么是选择的标准呢? 那就是这些被选的主成分所代表的主轴的长度之和占了主轴长度总和的大部分. 有些文献建议, 所选的主轴总长度占所有主轴长度之和的大约85%(也有的说80%左右)即

可. 其实, 这只是一个大体的说法, 具体选几个, 要看实际情况而定. 但如果所有涉及的变量都不那么相关, 就很难降维[1]. 不相关的变量就只有自己代表自己了. 用统计术语来说, 上述椭球的各个主轴的长短代表了该方向数据的方差, 而我们要寻找的是方差大(即数据变化大)的方向.

在引进主成分分析之前, 先看下面的例子.

例8.1 WHO数据(who.txt) 这是162个国家和地区的10个变量组成的数据, 数据摘自世界卫生组织的数据. 变量情况如下: x1: 青少年生育率(%)、x2: 人均国民收入, x3: 女小学生入学率(%)、x4: 男小学生入学率(%)、x5: 人口增长率(%)、x6: 城镇人口比率(%)、x7: 年龄中位数(%)、x8: 60岁以上比例(%), x9: 15岁以下比例(%)、x10: 每女性生育数. 目前的问题是, 能不能把感兴趣的10个变量用一两个综合变量来表示呢? 这一两个综合变量包含有多少原来的信息呢? 怎么解释它们呢? 能不能利用找到的综合变量来对国家和地区排序呢?

这一类数据所涉及的问题可以推广到对企业, 对学校进行分析、排序、判别和分类等问题. 这些在后面章节将会陆续引进. 下面首先介绍主成分分析.

例8.1的数据点是10维的, 也就是说, 每个观测值是10维空间中的一个点. 每一维代表了一个变量. 如果这些变量有些相关, 则可以把它们用某种综合变量来代表. 这就是一个降维的过程.

如何找主成分呢? 数学上是解数据相关阵的特征值问题, 下面的计算就是求该特征值的解. 对例8.1数据进行主成分分析, 通过R代码(包括输入数据)

```
w=read.table("who.txt",sep=",",header=T)
b=eigen(cor(w)) #解相关阵cor(w)的特征值问题
data.frame(b$va,b$va/sum(b$va),cumsum(b$va)/sum(b$va))
```

得到下面的输出:

主成分	特征值	特征值所占比例	特征值所占累积比例
1	6.7190	0.6719	0.6719
2	1.1536	0.1154	0.7873
3	0.8835	0.0884	0.8756
4	0.4674	0.0467	0.9223
5	0.4299	0.0430	0.9653
6	0.1703	0.0170	0.9824
7	0.1106	0.0111	0.9934
8	0.0336	0.0034	0.9968
9	0.0270	0.0027	0.9995
10	0.0052	0.0005	1.0000

[1]在所有变量都正交的情况下, 如果要达到一些文献所建议的85%主轴长度份额, 则必须选取85%的主成分, 似乎达到目的, 但毫无意义.

这里的特征值就是这里的10个主轴长度(相应方向的方差), 可以看出这10个特征值大小不一, 最大的有6.7190, 占主轴长度总和(或所有特征值的总和, 又叫总方差)的67.19%, 第二大特征值为1.1536, 占总方差的11.54%. 头两个主成分的特征值累积占了总方差的78.73%, 后面的特征值的贡献越来越少. 这可以从所谓悬崖碎石图(Scree Plot, 图8.2左图)看出. 图8.2右图为累积特征值比例. 由这些图表可以看出, 头两个特征值的确占了特征值总和的绝大部分. 因此, 选头两个主成分就可以了. 悬崖碎石图的名字意味着如果头一两个成分代表了大多数方差, 那么, 该图开始很陡, 其他分量就像悬崖落下的碎石一样基本靠近地面, 这也表示了选取主成分的一个原则, 即如果该图不陡, 那么主成分分析结果一定不好.

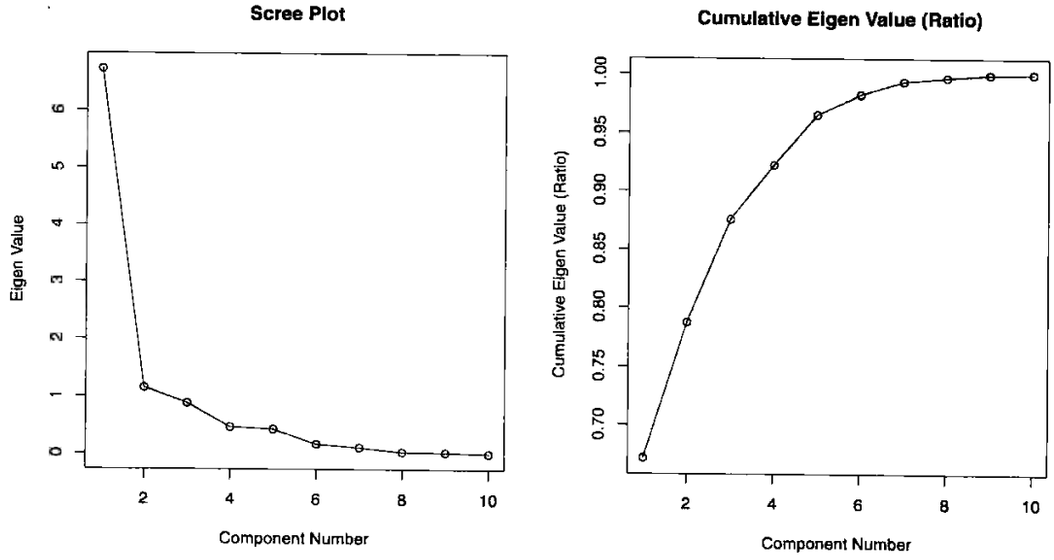

图 8.2 例8.1 十个成分的特征值的点图: 悬崖碎石图(左), 累积特征值比例(右).

图8.2是由下面代码绘的:

```
par(mfrow=c(1,2))
plot(b$va,type="o",main="Scree Plot",xlab="Component Number",
    ylab="Eigen Value")
plot(cumsum(b$va)/sum(b$va),type="o",
    main="Cumulative Eigen Value (Ratio)",
xlab="Component Number", ylab="Cumulative Eigen Value (Ratio)")
```

接下来的问题是怎么解释这两个主成分. 前面说过主成分是原数据十个变量的线性组合, 那么是怎么样的组合呢? 可以通过下面R语句输出所谓**载荷表(Component Matrix)**, 它也是特征向量, 只不过是单位乘以相应特征值的平方根, 因而有了变量和成分的相关系数的意义.

```
(loadings=sweep(b$ve,2,sqrt(b$va),"*"))
```

载荷表(里面数字为相应成分及变量之间的相关系数)

	成分									
变量	1	2	3	4	5	6	7	8	9	10
x1	0.80	-0.06	0.01	-0.57	-0.14	0.14	-0.08	0.00	-0.00	0.00
x2	-0.72	0.31	-0.48	-0.08	0.31	0.21	0.09	0.00	0.01	0.00
x3	-0.74	-0.64	-0.16	-0.06	0.01	-0.00	0.02	-0.10	-0.08	0.00
x4	-0.71	-0.66	-0.18	-0.10	0.06	-0.06	-0.00	0.10	0.07	-0.00
x5	0.80	0.00	-0.51	0.08	0.18	-0.10	-0.22	-0.03	0.02	0.00
x6	-0.66	0.28	-0.51	-0.04	-0.47	-0.10	0.03	0.00	-0.00	0.00
x7	-0.95	0.22	0.09	-0.11	0.09	-0.05	-0.10	0.03	-0.05	-0.05
x8	-0.86	0.24	0.19	-0.29	0.15	-0.20	0.01	-0.07	0.06	0.02
x9	0.97	-0.13	-0.13	-0.02	-0.01	-0.03	0.15	-0.06	0.06	-0.05
x10	0.92	0.07	-0.15	-0.16	0.17	-0.20	0.12	0.06	-0.08	0.01

上面表中的每一列[1]代表一个主成分, 作为原来变量线性组合的系数(比例). 比如第一主成分10个变量的线性组合系数为0.80, -0.72, -0.74, -0.71, 0.80, -0.66, -0.95, -0.86, 0.97, 0.92. 如果用$x_1, ..., x_{10}$表示原先的10个变量, 而用$y_1, ..., y_{10}$表示新的主成分, 那么, 第一和第二主成分y_1和y_2为:

$$y_1 = 0.80x_1 - 0.72x_2 - 0.74x_3 - 0.71x_4 + 0.80x_5 - 0.66x_6 - 0.95x_7 - 0.86x_8 + 0.97x_9 + 0.92x_{10}$$
$$y_2 = -0.06x_1 + 0.31x_2 - 0.64x_3 - 0.66x_4 + 0.00x_5 + 0.28x_6 + 0.22x_7 + 0.24x_8 - 0.13x_9 + 0.07x_{10}$$

这些系数称为主成分载荷(loading)[2], 它表示主成分和原先各变量的线性相关系数. 比如上面第一主成分y_1表示式中x_1的系数为0.80, 这就是说第一主成分和青少年生育率(x1)x_1的相关系数为0.80. 相关系数(绝对值)越大, 主成分对该变量的代表性也越大. 可以看得出, 第一主成分对各个变量解释得都很充分. 而最后的几个主成分和原先的变量就不那么相关了. 根据上面的公式, 可以对每个地区或国家根据各个变量(那10个原始变量的值)算出其主成分(比如y_1和y_2)的值, 称为**得分**或者**记分(score)**. 这样就可以按照这些主成分的大小对各个国家状况进行分析, 并利用主成分的意义来解释. 但是要注意的是, 利用这里的公式在计算每个观测值的主成分得分时应该对变量$x_1, ..., x_{10}$的数据列加以标准化, 当然, 计算机自动会做所有这些计算. 假定第j个变量的数据列为$x_{ij}, j = 1, ..., p$, 那么标准化的数据应该为

$$x^*_{ij} = \frac{x_{ij} - \bar{x}_{\cdot j}}{s_j}, \ i = 1, ..., n; \ j = 1, ..., p.$$

这里$\bar{x}_{\cdot j}$为第j列数据的样本均值, 而s_j为其样本标准差. 这样, 对于这个例子的第i个观测的头两个主成分得分(score)为

$$y_{i1} = 0.80x^*_{i1} - 0.72x^*_{i2} - 0.74x^*_{i3} - 0.71x^*_{i4} + 0.80x^*_{i5} - 0.66x^*_{i6} - 0.95x^*_{i7} - 0.86x^*_{i8} + 0.97x^*_{i9} + 0.92x^*_{i10}$$
$$y_{i2} = -0.06x^*_{i1} + 0.31x^*_{i2} - 0.64x^*_{i3} - 0.66x^*_{i4} + 0.00x^*_{i5} + 0.28x^*_{i6} + 0.22x^*_{i7} + 0.24x^*_{i8} - 0.13x^*_{i9} + 0.07x^*_{i10}$$

[1]这里的列向量分别是数据相关阵的各个特征值所相应的特征向量(eigenvector). 这里的向量不是单位向量, 而是单位特征向量乘以相应特征值的平方根(称为载荷). 载荷为对应的主成分和原先变量的相关系数. 有些文献(及软件)就用原始的单位特征向量的元素作为相应的主成分系数, 也称为载荷, 结果的主成分和这里的差一个大小等于相应特征值平方根的因子. 这种区别对于分析结果不会造成任何不同.

[2]单位特征向量也被会称为载荷, 但仅仅不是相关系数而已, 这不影响其他分析.

另外, 在不同软件中或不同的计算程序中, 对每个变量计算的得分列向量可能差一个常数.

为了更直观地解释主成分所代表的意义, 还能够把第一和第二主成分的载荷点出一个二维图以直观地显示它们如何解释原来的变量的. 这个图叫做载荷图(loading plot, 图8.3). 该图右面四个点是青少年生育率(x1)、人口增长率(x5)、15岁以下比例(x9)、每女性生育数(x10)四个点, 它们的坐标分别就是上面表中头两列的相应的行: x1: (0.80, −0.06), x5: (0.80, 0.00), x9: (0.97, −0.13), x10: (0.92, 0.07). 这说明第一主成分(横坐标)和这些变量正相关, 即第一主成分越大(正的值大), 则青少年生育率(x1)高, 人口增长率(x5)高, 15岁以下比例(x9)大, 每女性生育数(x10)多, 这是不发达国家的象征. 而图的左边则是其余的六个变量, 包括人均国民收入(x2)、女小学生入学率(x3)、男小学生入学率(x4)、城镇人口比率(x6)、年龄中位数(x7)、60岁以上人口比例(x8), 第一主成分和这些变量负相关, 即第一主成分小(负值大), 则人均国民收入(x2)高, 女小学生入学率(x3)高, 男小学生入学率(x4)高, 城镇人口比率(x6)高, 年龄中位数(x7)高, 60岁以上人口比例(x8)高, 这是发达国家的象征. 所以第一主成分的高低可以判断国家的发达程度, 也就是说, 第一主成分正方向越大, 国家综合起来越不发达. 第二主成分只代表了11.54%的信息, 远远没有第一主成分(代表67.19%的信息)那么显著, 但第二主成分的大小, 反映了教育状况的好坏, 因为只有小学入学率(x3和x4)是比较相关的变量(相关系数绝对值在0.6附近的仅有变量), 其他变量和第二主成分相关性不大. 第二主成分越大, 入学率越低(因为它和x3, x4负相关).

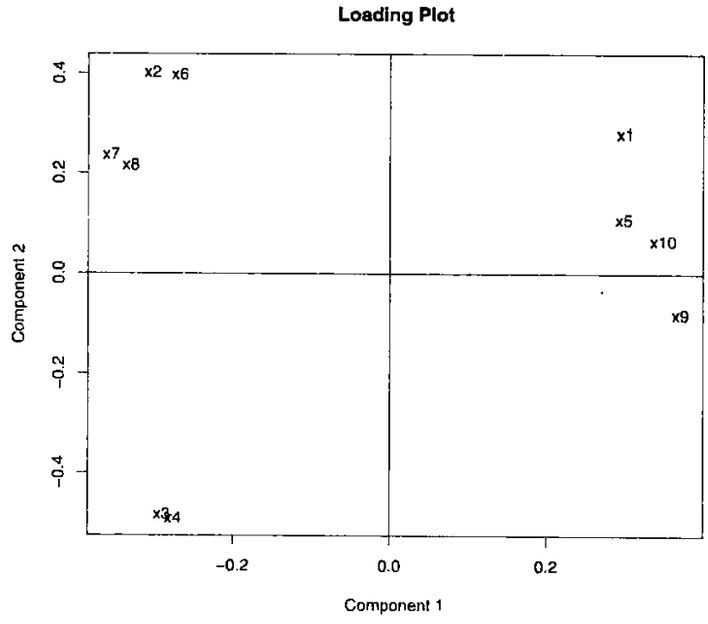

图 8.3 例8.1数据的10个变量的头两个主成分的载荷图, 显示了10个变量和这两个主成分的线性相关关系.

生成图8.3的代码为:

```
plot(b$ve[,1:2],type="n",main="Loading Plot",
xlab="Component 1",ylab="Component 2")
abline(h=0);abline(v=0);text(b$ve[,1:2],names(w))
```

要看各个国家在载荷图中的地位如何, 则可以点出它们的得分(图8.4), 前面已经给出了公式, 用R代码算并点出图来要简单得多:

```
w1=as.matrix(scale(w))
plot(w1%*%b$ve[,1:2],type="n",xlab="Comp 1",ylab="Comp 2")
text(w1%*%b$ve[,1:2],row.names(w),cex=0.5)
```

图8.4显示了各个国家的状况, 如在图8.3中说明的, 越靠近右边的国家越欠发达, 而左边的则发达. 这个图也印证了这种判断. 当然, 由于国家太多, 可能看不清楚, 读者可以自己生成这个图形来分析.

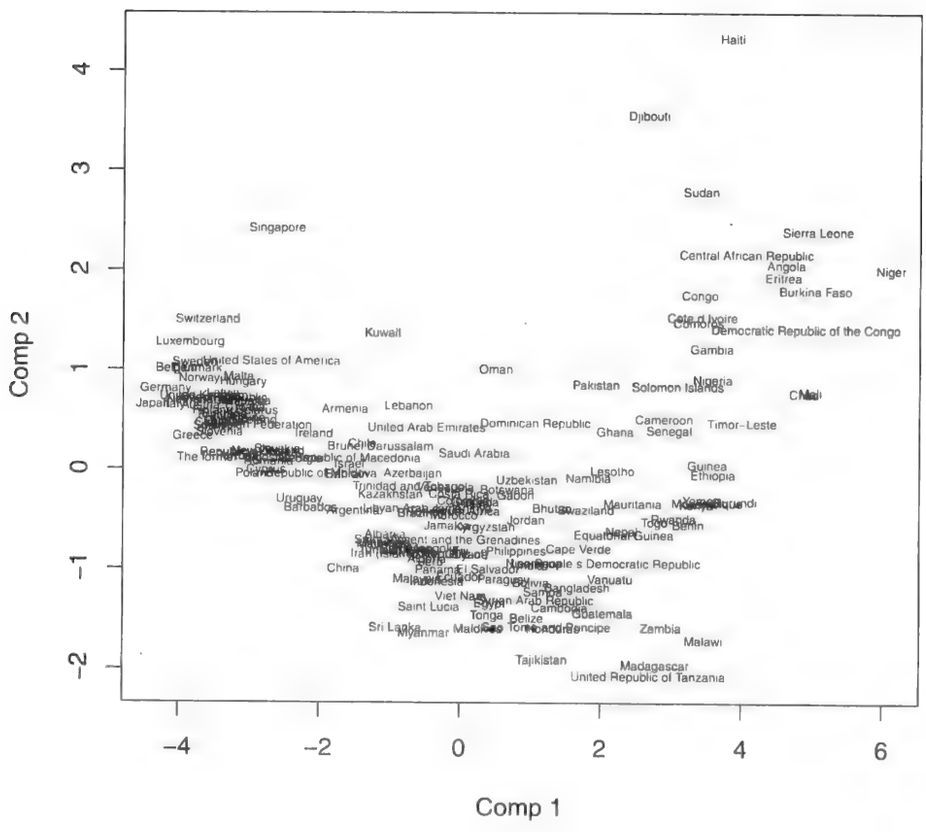

图 8.4　例8.1数据的各个国家相应于头两个主成分的得分.

实际上, 通过计算表明, 第一主成分最大(从大到小)的前10个国家为: Niger(尼日尔)、Sierra Leone(色拉里昂)、Burkina Faso(布基纳法索)、Mali(马里)、Democratic Republic of the Congo(刚果共和国)、Chad(乍

得)、Angola(安哥拉)、Eritrea (厄立特里亚)、Central African Republic(中非共和国)、Timor-Leste(东帝汶). 而第一主成分最小(从小到大)的前10个国家为: Japan(日本)、Germany(德国)、Italy(意大利)、Belgium(比利时)、Luxembourg(卢森堡)、Greece(希腊)、Sweden(瑞典)、Denmark(丹麦)、Norway(挪威)、Netherlands(荷兰). 从这个名单, 可以看出, 这种排名和仅仅用国民收入一项的排名不太一样.

思考一下:

1. "悬崖碎石图"形象地反映了选主成分的标准, 它意味着头一两个主成分很高, 而后面主成分就应该像从悬崖掉下的碎石一样很快降下来. 如果该图不很陡, 而形如一个缓斜坡, 说明降维效果不好. 不要为了凑够百分数而选取悬崖下面的"碎石".
2. 对互相正交的变量进行主成分分析, 得到的特征值均等于1, 这意味着所有成分同等重要.
3. 在变量多的时候, 有时不易解释主成分或因子.
4. 在不同的软件或程序中, 输出的载荷矩阵可能是单位特征向量, 也可能是单位特征向量乘以相应特征值的平方根, 而且也可能差一个正负符号, 它们对于各种分析没有任何影响(最多在载荷图中上下或左右调换位子), 但只有后者才代表主成分和相应变量的相关系数.

下面介绍的因子分析实际上是主成分分析的推广. 它和主成分分析的目的一致, 但分析更精密, 结果更有解释性.

8.1.2 因子分析

主成分分析从原理上是寻找椭球的所有主轴. 因此, 原先有几个变量, 就有几个主成分. 而因子分析是事先确定要找多少个成分(component), 这里称为**因子(factor)**(从数学模型本身来说必须事先确定因子个数, 但使用统计软件时, 或者使用者事先确定因子个数, 或者软件自动把符合某默认标准的因子都选入). 变量和因子个数的不同使得不仅在数学模型上, 而且在计算方法上, 因子分析和主成分分析有不少区别. 因子分析的计算要复杂一些. 根据因子分析模型的特点, 它还多一道工序: **因子旋转(factor rotation)**, 这个步骤可能会使结果更加满意. 当然, 对于计算机来说, 因子分析并不比主成分分析多费多少时间(可能多一两个选项罢了). 和主成分分析类似, 也可以根据相应特征值大小来选择因子的个数并展示初始的碎石图. 选择因子的标准也类似. 在输出的结果中, 因子分析也有**因子载荷(factor loading)**的概念, 代表了因子和原先变量的相关系数. 它也给出了二维载荷图, 其解释和主成分分析的载荷图类似.

还是以例8.1为例来看如何得到因子分析的结果. 利用得到下面包括输入数据的R代码(注意这里选项中标明只要2个因子):

```
w=read.table("who.txt",sep=",",header=T)
a=factanal(w,2,scores = "regression");a$loadings
```

用因子(成分)f_1和f_2来表示原来变量的关系(这些数字也称为载荷,):

```
       Factor1    Factor2
x1  -0.6604421 -0.3320026
x2   0.6897585  0.2105372
x3   0.3175013  0.9456340
x4   0.2964161  0.9204082
x5  -0.7311456 -0.3061111
x6   0.5654772  0.2217057
x7   0.9621764  0.2633326
x8   0.9189977  0.1949305
x9  -0.9212964 -0.3297795
x10 -0.7497948 -0.4757015
```

这个表说明10个变量和因子的关系. 为简单记, 用$x_1, ..., x_{10}$来表示那10个变量. 这样因子f_1和f_2与这些原变量之间的关系是(注意, 和主成分分析不同, 这里把成分(因子)写在方程的右边, 把原变量写在左边, 但相应的系数还是主成分和各个变量的线性相关系数, 也称为因子载荷):

$$
\begin{aligned}
x_1 &= -0.6604421 f_1 - 0.3320026 f_2 \\
x_2 &= \ 0.6897585 f_1 + 0.2105372 f_2 \\
x_3 &= \ 0.3175013 f_1 + 0.9456340 f_2 \\
x_4 &= \ 0.2964161 f_1 + 0.9204082 f_2 \\
x_5 &= -0.7311456 f_1 - 0.3061111 f_2 \\
x_6 &= \ 0.5654772 f_1 + 0.2217057 f_2 \\
x_7 &= \ 0.9621764 f_1 + 0.2633326 f_2 \\
x_8 &= \ 0.9189977 f_1 + 0.1949305 f_2 \\
x_9 &= -0.9212964 f_1 - 0.3297795 f_2 \\
x_{10} &= -0.7497948 f_1 - 0.4757015 f_2
\end{aligned}
$$

这里的系数所形成的散点图(也称载荷图, loading plot)直观地反映了这个特点(图8.5).

计算机还输出了这两个因子对方差的贡献: 各自贡献分别为51.4%, 24.8%, 而累积贡献为76.2%.

```
                Factor1 Factor2
SS loadings       5.136   2.481
Proportion Var    0.514   0.248
Cumulative Var    0.514   0.762
```

和主成分分析比较, 这里的第一因子和与教育有关的女小学生入学率(x3)、男小学生入学率(x4)已经不相关(相关系数分别为0.318和0.296, 但和其他变量的相关

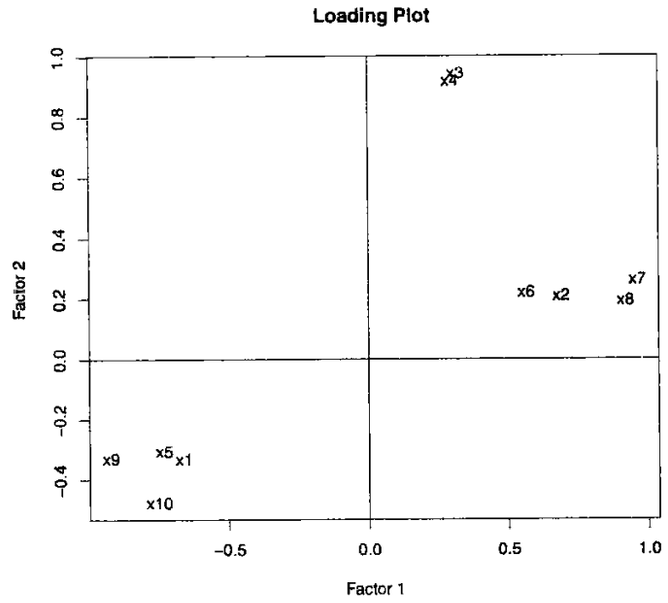

图 8.5 例8.1数据的头两个因子载荷图. 它和图8.3的区别为把与教育程度有关的两个变量完全从第一因子分离, 仅仅被第二因子代表. .

性都大于或接近0.6, 因此, 第一因子还是描述关于贫富、发达与不发达的特征, 与主成分分析载荷图相反, 越靠右(正的越大)的国家越发达, 左边而值越小(负的越大)的国家越不发达. 而第二因子和女小学生入学率(x3)与男小学生入学率(x4)很正相关, 相关系数分别为0.946和0.920, 也和生育率(x1)有点负相关. 因此, 第二因子主要描述教育普及程度. 第一因子可以起名为发达程度因子, 第二因子可以起名为教育因子. 从这个例子可以看出, 因子分析的结果比主成分分析解释性更强. 它把不同性质的变量区分得更清楚. 计算机输出中还有每个观测值在两个因子下的因子得分(score), 即对于162个国家($n = 162$)都算出两个得分($i = 1, 2$)

$$f_{ik} = \beta_{i1}x_{1k} + \cdots + \beta_{ip}x_{pk}, \quad i = 1, 2,\ p = 10,\ k = 1, ..., n.$$

根据这些得分, 可以画出类似于图8.4那样的得分图(图8.6). 这里的因子得分是用回归得到的(在前面因子分析代码中标明的选项).

和图8.4类似, 由于国家很多, 有的名字很长, 可能从图上看不清, 请读者自己重新产生这张图来分析. 图8.6是由下面代码画出的:

```
plot(a$scores,type="n",xlab="Factor 1",ylab="Factor 2")
text(a$scores,row.names(w),cex=0.5)
```

通过计算表明, 第一主因子最大(从大到小)的前10个国家为: Japan(日本)、Italy(意大利)、Germany(德国)、Croatia(克罗地亚)、Bulgaria(保加利亚)、Slovenia(斯洛文尼亚)、Switzerland(瑞士)、Latvia(拉脱维亚)、Belgium(比利时)、Finland(芬兰). 而第一主因子最小(从小到大)的前10个国家为: Malawi(马拉维)、Zambia(赞比亚)、United Republic of Tanzania(坦桑尼亚共和国)、Madagascar(马达加斯加)、Guatemala(危地马拉)、Sao

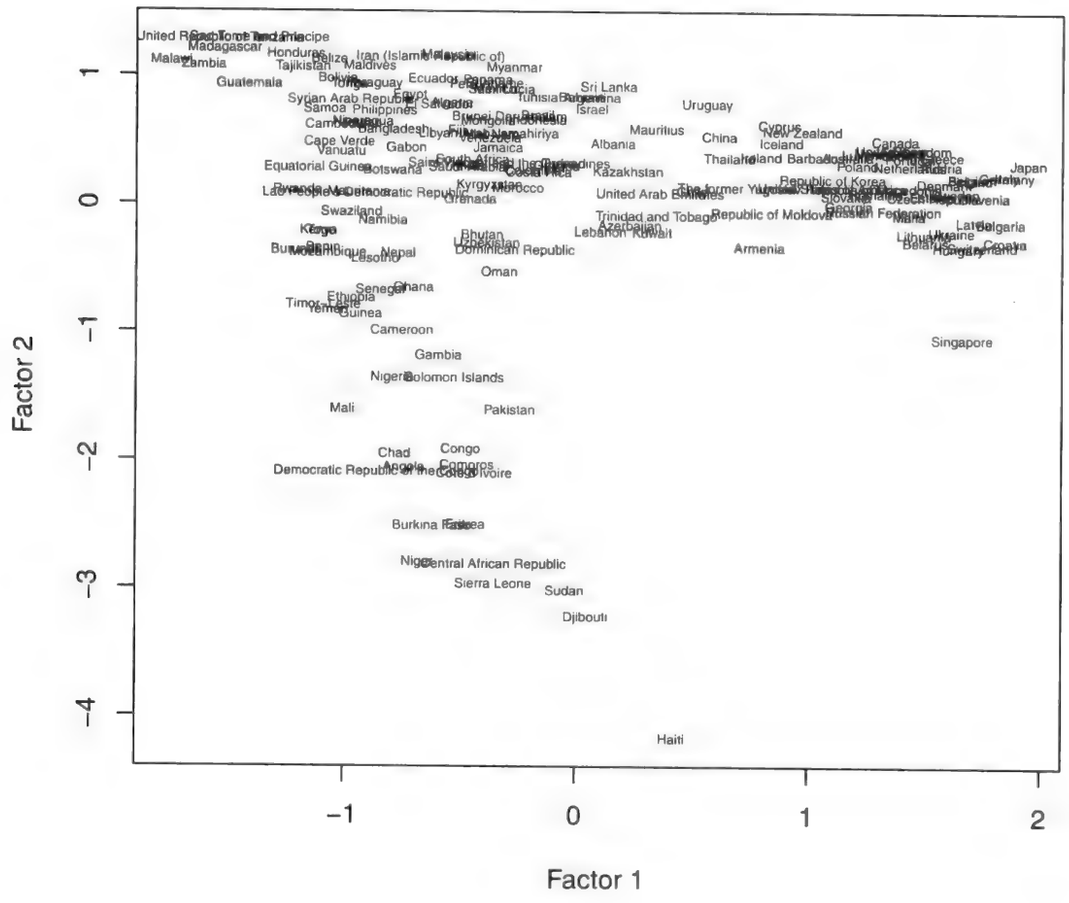

图 8.6 例8.1数据的各个国家相应于头两个因子的得分.

Tome and Principe(圣多美和普林西比)、Burundi(布隆迪)、Honduras(洪都拉斯)、Rwanda(卢旺达)、Tajikistan(塔吉克斯坦). 这个名单和主成分分析的第一主成分排名不尽相同, 这是各个成分和诸变量的相关(因而代表性)与因子和诸变量的相关(因而代表性)不一样所致.

思考一下:

1. 主成分分析和因子分析只能对互相相关的数量变量进行降维.
2. 如果变量没有近似的多维正态分布, 降维可能不理想.
3. 变量的选择很重要, 没有选入的变量, 绝对不会被主成分或因子所代表.
4. 因子分析载荷和得分的计算方法很多, 旋转方法也很多(这里用的是"最大方差法"), 因此不同软件、不同函数及不同选项算出来的结果不会完全相同.

8.1.3 因子分析和主成分分析的一些注意事项

可以看出, 因子分析和主成分分析都依赖于原始变量, 也只能反映原始变量的信息. 所以原始变量的选择很重要, 一定要符合进行分析所要达到的目标, 不能夹杂毫不相关的变量.

另外, 如果原始变量基本上互相独立, 那么降维就可能失败, 这是因为很难把很多独立变量用少数综合的变量概括. 数据越相关, 降维效果就越好. 那些选出的成分或因子代表了一些相关的信息(从相关性和线性组合的形式可以看出来). 根据这些信息可以帮助给这些成分或因子起合适的名字, 但并不总是可以给出满意的名字.

在得到分析的结果时, 并不一定会都得到如例8.1那样容易解释的清楚的结果. 这与问题的性质、选取的原始变量以及数据的质量等都有关系. 没有一个方法是万能的. 一个完美的世界就是由无数不完美的事物组成的.

在用因子得分进行排序时要特别小心, 特别是对于敏感问题. 由于原始变量不同, 因子的选取不同, 排序结果可以很不一样.

有人把主成分分析的特征向量按照特征根的大小的加权平均来得到所谓"综合指数", 这是没有道理的. 因为每个特征向量乘以任何实数之后还是特征向量, 都可以是主成分分析问题的解, 不同软件及不同选项得到的结果并不一样. 即使采取单位特征向量还有可能差一个正负号. 假定数据有10个变量, 可以得到10个主成分, 那么, 考虑到符号变化的所有可能, 从该数据一共可以得到$2^{10} = 1024$种不同的"综合指数", 这不是很荒谬的事情吗? 此外, 主成分是互相正交的, 把正交变量"加权平均"更是不可思议和难以解释. 通常, 得到综合指数的目的是为了容易解释, 但是一个主成分的大小往往代表多层意义(正负面都可能同时存在). 假定从一个数据的若干变量中选出两个主成分, 第一主成分的数值大(正数)代表效率, 第一主成分小(负值大)代表公平; 而第二主成分大(正数)代表GDP高, 第二主成分小(负值大)代表腐败. 那么无论如何在$2^2 = 4$种"综合指数"中选哪一个都无法解释. 显然这四种选择的解释为1. 效率高、不公平、GDP高、腐败; 2. 效率低、公平、GDP高、腐败; 3. 效率高、不公平、GDP低、清廉; 4. 效率低、公平、GDP低、清廉. 难道这些"综合指数"的选择对使用指数的人方便吗? 当然, 人们可以用任何方法编制一些符合他们需要的"综合指数", 但不要以主成份分析的"科学性"作为理由.

8.2 把对象分类: 聚类分析

俗语说, 物以类聚、人以群分. 但什么是分类的根据呢? 比如, 要想把中国的县分成若干类, 就有很多种分类法, 可以按照自然条件来分, 比如考虑降水、土壤、植被、日照、湿度等各方面. 也可以考虑收入、教育水准、医疗条件、基础设施等指标. 既可以用某一项来分类, 也可以同时考虑多项指标来分类.

对于一个数据, 人们既可以按照观测值(行)对变量(指标)进行分类(相当于对数据中的列分类), 也可以按照变量(列)对观测值(事件, 样品)来分类(相当于对数

据中的行分类). 比如利用上一章例8.1的数据就可以对国家或地区按照经济、人口、教育等分类. 当然, 并不一定事先假定有多少类, 完全可以按照数据本身的规律来分类. 本章要介绍的分类的方法称为**聚类分析(cluster analysis)**. 有人称按照观测值对变量的分类为R型聚类, 而称按照变量对观测值的分类称为Q型聚类. 其实无所谓, 这两种聚类在数学上是对称的, 没有什么不同.

例8.2　交通数据(trans.txt) 该数据收集了一些国家和地区的运输数据, 变量包括机场数目(Airports)、铁路公里数(Railways.km)、公路公里数(Roadways.km)、水路公里数(Waterways.km)、商船数(Merchant.marine)等5个变量. 而国家或地区的名字(Country.Area)则在行名字中. 现在希望利用这5个变量来分类. 如果按照这5个指标的任何一项来分类, 问题就很简单了, 只要把该指标相近的点放到一起就行了. 如何同时根据这5个变量来聚类呢? 其想法也类似, 就是把距离近的放到一起. 这样就出现下面要提到的距离的定义和度量等问题. 该数据有80个观测值.

8.2.1　如何度量距离远近?

对例8.2数据的最简单的分类就是对一项指标(比如机场数)进行分类, 这些数值在直线上形成许多点. 这样就可以把直线上距离近的点放到一起. 如果再加上一个变量, 比如公路里程, 那么, 这两个变量就形成二维平面上的一些点, 也可以按照平面上的距离远近来分类. 三维或者更高维的情况也是类似, 只不过三维以上的图形无法直观地画出来而已.

在例8.2的数据中, 每个观测都有5个变量值. 这就是5维空间点的问题了. 按照远近程度来聚类需要明确两个概念: 一个是点和点之间的距离, 一个是类和类之间的距离. **点间距离**有很多定义方式. 最简单的是熟知的欧氏距离. 根据距离来决定两点间的远近是最自然不过了. 当然还有一些和距离不同但起类似作用的概念, 比如相似性等, 两点越相似, 就相当于距离越近.

由一个点组成的类是最基本的类, 如果每一类都由一个点组成, 那么点间的距离就是**类间距离**. 但是如果某一类包含不止一个点, 那么就要确定类间距离. 类间距离是基于点间距离定义的, 它也有许多定义的方法, 比如两类之间最近点之间的距离可以作为这两类之间的距离, 也可以用两类中最远点之间的距离作为这两类之间的距离, 当然也可以用各类的中心之间的距离来作为类间距离. 在计算时, 各种点间距离和类间距离的选择是通过统计软件的选项实现的. 不同的选择的结果可能会不同.

有了上面的点间距离和类间距离的概念, 就可以介绍聚类的方法了. 这里介绍两个简单的方法.

8.2.2　事先要确定分多少类: k均值聚类

前面说过, 聚类可以走着瞧, 不一定事先确定有多少类, 但是这里的**k均值聚类(k-means cluster**, 也叫**快速聚类, quick cluster)** 却要求你先说好要分多

少类. 看起来有些主观.[1] 假定你说分3类, 一些软件还给你选择三个点作为"聚类种子"的机会, 如果你不选, 那软件可以随机为你选种子. 计算中, 把这3个点作为三类中每一类的基石. 然后, 根据与这三个点的距离远近, 把所有点分成三类. 再把这三类的中心(均值)作为新的基石或种子(原来的"种子"就没用了), 重新按照距离分类. 如此迭代下去, 直到达到停止迭代的要求(比如, 各类最后变化不大了, 或者迭代次数太多了). 如果客观上各类很容易区分, 聚类种子的选择并不必太认真, 它们很可能最后还会分到同一类中. 但如果各类区别不明显, 对于不同种子的选择有可能导致聚类结果不同, 因此对同一个数据重复计算时, 如果随机选定的种子不一样, 结果可能有异. 下面用例8.2的数据来描述k均值聚类.

就例8.2来说, 假定要把这些国家或地区按照5个关于运输的变量分成5类. 利用R语句(包括读入数据)

```
w=read.table("trans.txt",header=T)
set.seed(44);a=kmeans(w,5)
```

可以得到最后的5类的中心(在5维空间中的)坐标:

```
  Airports Railways.km Roadways.km Waterways.km Merchant.marine
1  1692.0000    55984.67  2332954.67    62833.333         821.000
2   150.4746     3405.39    73812.12     2779.593         225.661
3 14947.0000   226612.00  6465799.00    41009.000         422.000
4   488.9000    14071.70   359264.40     7440.000         344.800
5   631.4286    41401.14   883353.57    17624.714         381.571
```

由R语句`for(i in 1:5)print(row.names(w)[a$clus==i])`得到这5类国家为(有一类国家太多, 略去):

类别	国家或地区
1	Brazil, China, India
2	59个国家和地区(名称略去)
3	United States
4	Argentina, Bangladesh, Indonesia, Italy, Mexico, Poland, Sweden, Turkey, United Kingdom, Vietnam
5	Australia, Canada, France, Germany, Japan, Russia, Spain

[1] 也有自动寻求最优k值的聚类软件, 这里不予介绍.

思考一下:

1. 根据你对这些国家或地区的了解, 讨论上面根据运输数据聚类分划的5个类别有没有道理.
2. 如果把例8.2数据的变量除以各个国家和地区的人口, 结果会有多大不同?
3. 试着用例8.2数据通过k均值聚类, 把观测值分成6类、7类等.
4. 关于k均值聚类究竟应该分成多少类有很多讨论, 产生了不少算法. 比如, 把类间距离的平方和与类内距离平方和之比达到最大为标准来确定k值. 对于下面的分层聚类, 也有类似的问题.

8.2.3 事先不用确定分多少类: 分层聚类

另一种聚类称为**分层聚类**或**系统聚类(hierarchical cluster)**. 开始时, 有多少点就是多少类. 它第一步先把最近的两类(点)合并成一类, 然后再把剩下的最近的两类合并成一类, 这样下去, 每次都少一类, 直到最后只有一大类为止. 显然, 越是后来合并的类, 距离就越远.

继续对例8.2数据进行聚类分析. 为了演示清楚, 把上面用k均值聚类得到的包括59个国家那一类从数据中去掉, 对剩下的21个国家进行分层聚类, 用R语句

```
w1=w[a$clus!=2,];hh=hclust(dist(w1), "ave")
plot(hh,labels=row.names(w1) ,xlab="Country or Area")
```

得到图8.7的结果(**聚类树形图, dendrogram**). 在图中, 纵向的尺度是和计算出来的距离成比例的, 因此, 可以直观地看出各个类别的远近.

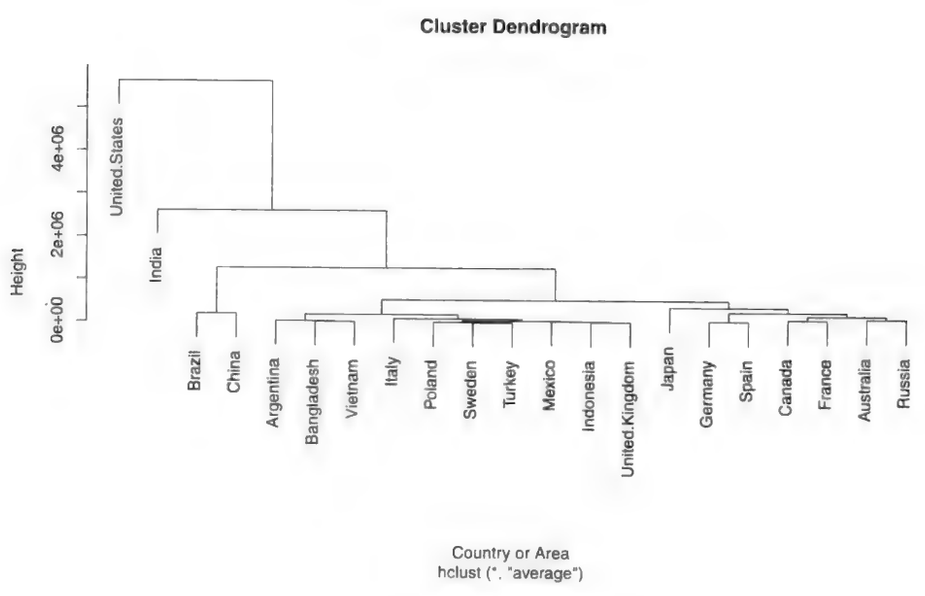

图 8.7　例8.2数据用分层聚类法对去掉一些观测值后的例8.2数据所进行的分层聚类.

可以看出, 如果要分成两类, 则在右边只有两条纵线处横向“切开”, 得到美国为一类, 其他国家为另一类, 如果要分成三类, 则在只有三条纵线处“切开”, 得到美国为一类, 印度为一类, 其余的为第三类.

例8.3 大城市建筑数据(cities0.txt) 这是世界一些大城市的建筑数据, 包括人口、面积(km^2)、高层建筑数目、高层建筑的点数(按照每个建筑的层数确定的该城市建筑的总点数). 根据这个数据, 用分层聚类法, 把城市分类, 包括读入数据及手工选择4类的代码如下:

```
w=read.table("cities0.txt",sep=",",header=T)
hh=hclust(dist(w), "ave");
plot(hh,labels=row.names(w),cex=0.8);a=identify(hh)
```

树形图显示在图8.8中.

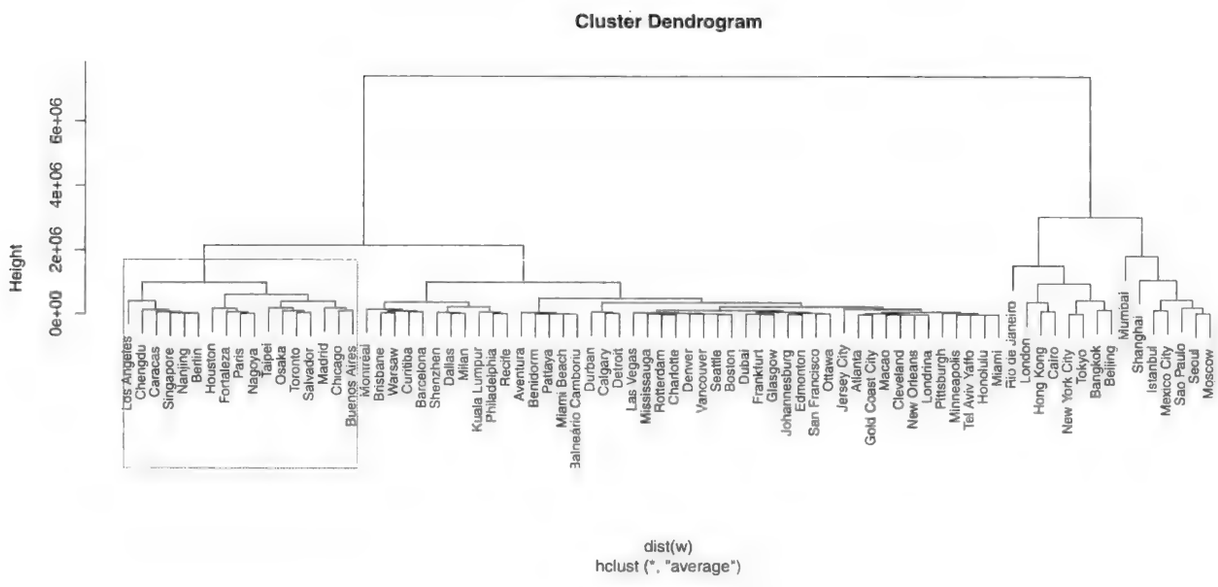

图 8.8　用分层聚类法对例8.3数据所进行的分层聚类.

这里尝试把这些城市分成4类, 由于树形图字小, 读者可自行产生这个图形.

思考一下:

1. 人们在回归中经常发现数据中的点不一定能够用一个回归模型来描述, 这时, 可以用聚类方法把数据中的观测值分成一些子群, 再分别进行回归, 或者把类别当成一个定性变量, 和其他变量一起进行有交互作用的回归.

2. 分层聚类中对点间距离和类间距离的不同选择可能会产生不同的聚类结果.

8.2.4 聚类要注意的问题

显然, 聚类结果主要受所选择的变量影响. 如果去掉一些变量, 或者增加一些变量, 结果会很不同. 相比之下, 聚类方法的选择则没有变量选择那么重要. 因此, 聚类之前一定要目标明确. 比方说, 如果在例8.2的国家或地区根据交通分类的问题再加上涉及地理、经济及人口等信息的变量, 得到的结果就可能很不一样了.

另外就分成多少类来说, 也要有道理. 只要你高兴, 无论从k均值聚类或分层聚类都可以得到任何可能数量的类. 但是, 聚类的目的是要使各类之间的距离尽可能地远, 而类中点之间的距离尽可能地近, 而且分类结果还要有令人信服的解释. 虽然使用数学可以定义一些准则, 但最主要的是, 一定要搞清聚类的原始动机和目的.

8.3 两组变量之间的相关: 典型相关分析

8.3.1 两组变量的相关问题

前面第七章已经介绍了如何衡量两个变量之间是否相关的问题, 这是用简单的公式就可以解决的问题. 但是, 如果有两组而不是两个变量, 如何能够表明它们之间的关系呢? 下面看一个例子(例8.4).

例8.4 电视打分数据(tv.txt) 业内人士和观众对于一些电视节目的观点有什么样的关系呢? 下面数据(tv.txt)是不同的人群对30个电视节目所作的平均评分. 观众评分来自低学历(led)、高学历(hed)和网络(net)调查三种, 它们形成第一组变量, 而业内人士的评分来自包括演员和导演在内的艺术家(arti)、发行(com)与业内各部门主管(man)三种, 形成第二组变量. 人们对这样两组变量之间的关系感到兴趣.

对电视节目的打分

No.	led	hed	net	arti	com	man	No.	led.1	hed	net	arti	com	man
1	86	43	85	43	93	71	16	39	80	71	76	52	81
2	99	74	99	78	99	89	17	65	5	53	11	67	41
3	37	22	10	27	24	33	18	28	11	31	12	23	35
4	5	19	56	13	11	38	19	50	32	68	23	49	58
5	45	43	55	39	54	58	20	69	98	69	97	81	99
6	21	32	21	34	35	32	21	55	99	78	97	60	90
7	36	78	48	75	42	78	22	36	11	5	15	26	5
8	69	31	85	32	70	52	23	77	18	61	27	68	54
9	40	98	36	99	64	86	24	67	33	95	34	59	61
10	26	14	40	8	25	21	25	45	87	46	85	67	80
11	51	68	38	68	48	72	26	61	72	63	63	62	75
12	63	86	79	87	76	95	27	41	63	74	55	50	76
13	39	80	57	80	55	68	28	6	5	13	5	5	13
14	78	40	72	42	75	58	29	28	53	35	51	31	59
15	56	49	54	48	52	61	30	66	20	79	18	67	55

如果对这6个变量进行两两相关分析, 可以得到15个相关系数, 但是从这些相关系数中很难得到这两组(每组有3个变量)变量之间的关系. 如果能把每一组变量用一个变量来代表, 那么, 多个变量与多个变量之间的相关就可以化为两个变量之

间的相关, 问题也就简单了. 人们可能会说, 可以用主成分分析, 各自找到各自的代表, 再看它们之间的相关不就行了? 但是, 各自的主成分只能代表自己, 很可能互相不相关, 因此, 一定要找出可以联系两组变量的代表. 这有些像两个完全不同的国家之间的谈判, 谈判代表不能完全只代表各自的利益, 而要能够在两国之间基于它们的共同点来建立联系.

这一章的目的就是为人们感兴趣的两组变量各找到一个(或多个)有综合意义的代表变量, 并通过研究这两个代表变量之间的关系来考察两组变量之间的关系.

8.3.2 典型相关分析

现在要为每一组变量选取一个综合变量作为代表, 而一组变量最简单的综合形式就是该组变量的线性组合. 由于一组变量可以有无数种线性组合(线性组合由相应的系数确定), 因此必须找到既有意义又可以确定的线性组合. **典型相关分析(canonical correlation analysis)**就是要找到这两组变量线性组合的系数使得这两个由线性组合生成的变量(和其他线性组合相比)之间的相关系数最大.

假定两组原始变量为$X_1, ..., X_p$和$Y_1, ..., Y_q$, 而需要寻找用来代表这两组变量的新综合变量V和W称为**典型变量(canonical variable)**, 由下面的表达式给出(其中系数$a_1, a_2, ..., a_p$及$b_1, b_2, ..., b_q$是在典型相关分析中想要得到的):

$$V = a_1X_1 + \cdots + a_pX_p$$
$$W = b_1Y_1 + \cdots + b_qY_q$$

那么, 典型相关分析的问题就在于要寻找系数$a_1, a_2, ..., a_p$及$b_1, b_2, ..., b_q$, 使得典型变量V和W之间的相关关系最大.

这种相关关系是用**典型相关系数(canonical correlation coefficient)**来衡量的. 这里所涉及的主要的数学工具还是矩阵的特征值和特征向量问题(见本章后面的公式), 而所得的特征值与V和W的典型相关系数有直接联系. 由于特征值问题的特点, 实际上找到的是多组典型变量$(V_1, W_1), (V_2, W_2), ...$, 其中$V_1$和$W_1$最相关, 而$V_2$和$W_2$次之等等, 而且$V_1, V_2, V_3, ...$之间及$W_1, W_2, W3, ...$之间互不相关. 这样又出现了选择多少组典型变量$(V, W)$的问题了. 这其实很简单, 就像在主成分分析中选主成分一样, 只要选择特征值累积总贡献占主要部分的那些即可. 当然, 软件还会输出一些检验结果, 于是只要选择显著的那些(V, W). 对于实际问题, 还要看选取的(V, W)是否有意义, 是否能够说明问题才行. 至于得到(V, W)系数的计算, 则很简单, 下面就例8.4的数据进行分析.

下面利用软件包CCA[1]实行这两组变量的典型相关分析. 为了方便把前三个变量(led, hed, net)的一组命名为X, 另一组(arti, com, man)命名为Y. 进行典型相关分析的R代码如下:

```
w=read.table("tv.txt",header=T)
X=w[,1:3];Y=w[,4:6]
```

[1]Ignacio González and Sébastien Déjean (2009). CCA: Canonical correlation analysis. R package version 1.2. http://CRAN.R-project.org/package=CCA.

```
library(CCA)
(res=cc(X,Y))
```

由于两组中变量个数均为3, 因此最多有3对典型相关变量. 一般来说如果两组变量个数不一样, 比如$p > q$, 则典型相关变量个数不会超过q. 输出中包括了这3个典型相关系数: 0.9954405 0.9528195 0.6373226 (用res$cor来显示) 这3对典型相关变量为各自组中变量的线性组合, 可用代码res$xcoef和res$ycoef得到其系数的输出(见下面). 注意输出中X组的典型相关变量称为X, 而Y组的典型相关变量称为Y[1], 在其他软件中通常也是把一组称为"因变量", 另一组称为"自变量", 但实际上它们是对称的, 使用这种称谓仅仅是为了方便.

```
$xcoef
             [,1]         [,2]         [,3]
led -0.006674773 -0.03523045  0.054341051
hed -0.031823575  0.01247933  0.005196029
net  0.002099295 -0.01257811 -0.059215023

$ycoef
               [,1]         [,2]         [,3]
arti -0.0286177622  0.03040737  0.06616270
com  -0.0008426431 -0.04568546  0.04865502
man  -0.0060022012 -0.01391857 -0.11696518
```

还可以从此得到各个观测值的得分(即xscores和yscores), 用代码res$scores$xscores和res$scores$yscores显示(由于太长, 这里不显示). 这些得分类似于因子分析中的因子得分, 为各个观测值与典型变量做同样的线性组合所得. 下面是输出中的由典型相关变量得到的(X组的)x得分x-score和(Y组的)y得分y-score分别与X组和Y组原始变量(即我们的变量代码X和Y组)之间的相关系数(产生四个表, 对应于四种相关组合):

```
$scores$corr.X.xscores
          [,1]       [,2]        [,3]
led -0.3325178 -0.9248417  0.18466107
hed -0.9932899  0.1008356 -0.05663309
net -0.3826908 -0.7530492 -0.53522395

$scores$corr.Y.xscores
           [,1]        [,2]        [,3]
arti -0.9924136  0.06162569  0.02770038
com  -0.5684258 -0.77295364  0.08019883
```

[1]我们在程序中把一组命名为X另一组为Y, 是为了和这个R函数输出一致. 无论我们选不选名字, 或者选择什么名字, 该函数输出中还是用X和Y, 其他软件也类似, 比如SPSS在输出中, 总是把一组成为因变量, 另一组称为自变量.

```
man  -0.9180073 -0.26087863 -0.17402564

$scores$corr.X.yscores
          [,1]        [,2]        [,3]
led -0.3310017 -0.88120725  0.11768867
hed -0.9887610  0.09607813 -0.03609355
net -0.3809460 -0.71751994 -0.34111031

$scores$corr.Y.yscores
           [,1]        [,2]        [,3]
arti -0.9969593  0.06467719  0.04346368
com  -0.5710294 -0.81122775  0.12583711
man  -0.9222121 -0.27379648 -0.27305738
```

上面4个表也可以分别用下面4行代码获得:

```
res$scores$corr.X.xscores
res$scores$corr.Y.xscores
res$scores$corr.X.yscores
res$scores$corr.Y.yscores
```

如何解读这些内容呢? 下面举例说明:

(1) 从第一个表可第一列以看出, V_1和高学历的人(hed)的相关系数为-0.9932899, 和另外两个不相关, 因此V_1只与高学历的观点有关; 从第四个表第一列可以看出W_1与艺术家(arti)及主管(man)相关(相关系数分别为-0.9969593和-0.9222121); 而V_1和W_1为最相关的一对典型变量, 这说明, 高学历的与艺术家及主管观点较一致.

(2) 从第一个表第二列可以看出, V_2和低学历的人(led)与网民(net)的相关系数较高, 分别为-0.9248417和-0.7530492, 因此V_2与低学历和网民的观点有关; 从第四个表第二列可以看出W_2只与发行(com)相关(相关系数为-0.81122775); 而V_2和W_2为第二相关的一对典型变量, 这说明, 低学历及网民与发行观点较一致.

(3) 注意, 上面四个表有一些信息有些重合. 第一和第三, 第二和第四个表有些类似. 这是必然的, 以第二表为例, 第二表第一列显示了V_1和高学历(hed)及主管(man) 很相关(相关系数分别为-0.9924136及-0.9180073), 这和上面(1)的结论符合.

上面仅仅列出了统计结果, 到底如何从对传媒的理解来解释这些结果, 则留给读者.

8.4　列联表行变量和列变量的关系: 对应分析

在因子分析中, 或者对变量(列中的变量)进行分析, 或者对样品(观测值或行中的变量)进行分析, 而且常常把每一种分析结果画出载荷图来看各个变量之间的接近程度. 典型相关分析也只研究列中两组变量之间的关系. 然而, 在很多情况下, 所关心的不仅仅是行变量自身之间或列变量自身之间的关系, 而是若干行变量与若干列变量之间的关系, 或者是列联表中行变量和列变量的各个水平之间的相互关系, 这是因子分析等方法所没有涉及的. 这里介绍的用图来描述列联表行列变量之间关系的方法, 称为**对应分析(correspondence analysis)**方法.

这里我们用例7.8的眼睛和头发颜色数据(HEColor.txt)中的头发及眼睛颜色的列联表:

	Eye			
Hair	Blue	Brown	Green	Hazel
Black	20	68	5	15
Blond	94	7	16	10
Brown	84	119	29	54
Red	17	26	14	14

人们可以对这个列联表进行前面所说的χ^2检验来考察行变量和列变量是否独立. 前面已经知道这个检验很显著: p值等于2.2×10^{-16}. 看来两个变量的确不独立. 但是如何用像因子分析的载荷图那样的直观方法来展示这两个变量各个水平之间的关系呢? 这就是本章要介绍的对应分析方法内容, 它被普遍认为是探索性数据分析的范畴, 读者只要能够会用数据画出描述性的点图, 并能够理解图中包含的信息即可. 对应分析还可以描述多于二维的数据, 但由于多维的图形展示不那么容易看懂, 这里不予介绍.

在对应分析中. 可以找到行和列的若干有意义的代表, 分别称为**行得分(row score)**和**列得分(column score)**, 它们互为对方的加权均值, 而且它们之间有不同程度的相关. 这有些像典型相关分析, 只不过那里是两组列变量, 而这里是行变量和列变量. 这些概念的数学意义会在后面小结中给出. 为了得到最直观的叠加的二维散点图, 一般选择两对行列得分(最多不超过三维). 选择的维数的代表性主要看它们之间的相关程度, 选取相关系数(也称为典型相关系数)最大的两个.

下面通过对例7.8数据的计算和结果分析来介绍对应分析.

首先看例7.8数据的对应分析的一个主要结果, 即图8.9, 这个图是用下面的R代码实现的(包括数据输入):

```
w=read.table("HEcolor.txt",header=T)
w1=xtabs(Freq~Hair+Eye,w)
library(MASS);(a=corresp(w1, nf=2))
biplot(a,xlim=c(-1,1))
```

注意, 这里的数学主要是解矩阵的特征值和特征向量问题, 由于数学上, 一个特征向量乘以任何正负实数之后还是特征向量, 所以各个软件所画出的图可能会方向相反, 尺度也可能有区别, 但对应分析的结果, 即各个变量之间的关系是不会因此而显示出不同.

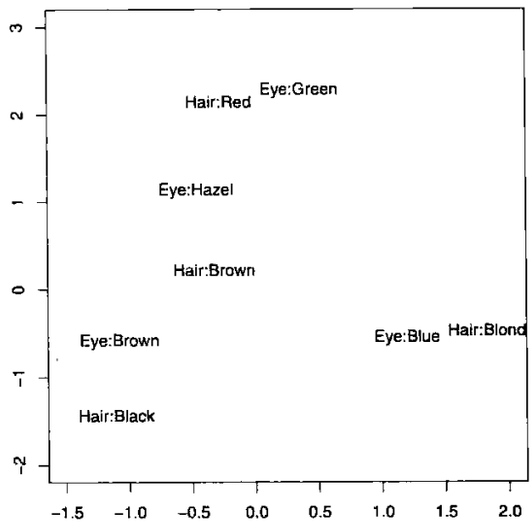

图 8.9 对于头发颜色和眼睛颜色关系(例7.8)的对应分析点图.

除了图8.9之外, 上面代码还计算并输出了两对行列变量的典型相关系数(第一个为0.4569165为第二个的3倍多)以及行得分和列积分,它们是图7.9中8个点的坐标:

```
First canonical correlation(s): 0.4569165 0.1490859

 Hair scores:
            [,1]       [,2]
Black -1.1042772 -1.4409170
Blond  1.8282287 -0.4667063
Brown -0.3244635  0.2191109
Red   -0.2834725  2.1440145

 Eye scores:
            [,1]       [,2]
Blue   1.1980612 -0.5564193
Brown -1.0771283 -0.5924202
Green  0.3540108  2.2741218
Hazel -0.4652862  1.1227826
```

图8.9体现了头发颜色和眼睛颜色的关系, 主要看横坐标的相对位置, 这是因为第一对变量的典型相关系数比第二对大得多. 显然, 金发(Blond)和蓝眼(Blue)很相关, 黑头发(Black)和棕色眼睛(Brown)较接近. 这和遗传学的结论是一致的. 该图直观地展示了前面单独用χ^2检验所无法看出的关系.

8.5 小结

8.5.1 本章的概括和公式

1. 主成分分析和因子分析

主成分分析和因子分析的主要目的就是用少数几个互相正交的变量(因子分析也可以选择不正交的因子)来代表原始数据中较多的相关的变量. 这些新变量叫做因子或者成分. 因子分析和主成分分析会产生出因子载荷和因子得分. 因子载荷代表了每个因子(成分)与原先每个变量的线性相关系数, 可以用之对因子(成分)进行解释(甚至命名). 因子得分用原来变量的线性组合来表示每个因子. 在主成分分析中, 选择成分的标准是根据各个成分方差大小来决定的. 方差就是数据相关阵的特征值, 对应于数据相关阵的特征向量一般则称为载荷, 但代表成分和变量之间相关系数的载荷在数值上等于该特征值的平方根乘以与之对应的单位特征向量. 如果所选成分的累积方差和总方差之比很显著, 那么就不再选更多的成分了. 在因子分析中, 也可以按照主成分分析选成分的方法来选择因子, 但也有其他方法.

按照数学原理, 主成分为数据相关阵的特征向量, 而每个成分的方差为相应的特征值. 记特征向量为$a_i = (a_{i1}, ..., a_{ip})^T$(假定数据有$p$个变量), 而相应的特征值记为$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, 则主成分分析的成分$y_i$和原来变量$x_i$之间的关系为:

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p \\ &\vdots \\ y_p &= a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pp}x_p \end{aligned}$$

一般来说, 软件输出特征向量为单位向量, 也有的输出是单位向量乘上相应特征值的平方根$\sqrt{\lambda_i}$. 如果a_i是单位特征向量乘了$\sqrt{\lambda_i}$之后的向量, 那么, a_{ij}为第i个成分y_i和第j个原先的变量x_j之间的线性相关系数. 无论输出的是什么, 由于特征向量乘以一个常数(无论正负)都还是特征向量, 所以, 后续分析除了是否是相关系数之外不会由于特征向量差个常数因子而有所区别, 而且, 特征向量习惯上被称为载荷(无论是否是相关系数). 头两个主成分的载荷图就是下面坐标的点组成的: $(a_{11}, a_{21}), (a_{12}, a_{22}), ..., (a_{1p}, a_{2p})$.

因子分析的因子f_i和原来变量x_i之间的模型和关系(假定原先有p个变量及需要求m个因子, $m < p$). 因子分析的理论模型为

$$x_1 - \mu = a_{11}f_1 + a_{12}f_2 + \cdots + a_{1m}f_m$$

$$
\begin{aligned}
x_2 - \mu &= a_{21}f_1 + a_{22}f_2 + \cdots + a_{2m}f_m \\
&\vdots \\
x_p - \mu &= a_{p1}f_1 + a_{p2}f_2 + \cdots + a_{pm}f_m
\end{aligned}
$$

这里的a_{ij}为第i个变量x_i和第j个因子f_j之间的线性相关系数, 也称为载荷. 由于这个模型复杂, 理论分析也复杂. 这里不做详细讨论. 头两个因子的载荷图就是下面坐标的点组成的: $(a_{11}, a_{12}), (a_{21}, a_{22}), ..., (a_{p1}, a_{p2})$. $\sum_{j=1}^{m} a_{ij}^2$称为**共性方差** (也称**公共方差或变量共同度, common variance, communalities**).

从数据经过因子分析得到的因子得分函数为:

$$
\begin{aligned}
f_1 &= \beta_{11}x_1 + \beta_{12}x_2 + \cdots + \beta_{1p}x_p \\
f_2 &= \beta_{21}x_1 + \beta_{22}x_2 + \cdots + \beta_{2p}x_p \\
&\vdots \\
f_m &= \beta_{m1}x_1 + \beta_{m2}x_2 + \cdots + \beta_{mp}x_p
\end{aligned}
$$

由于每个观测值都有p个数: $x_1, x_2, ..., x_p$, 所以可以按照因子得分函数算出所有观测值的因子得分.

2. 聚类分析

本章介绍了两种聚类方法. 聚类是基于距离这个概念的. 首先要定义两点之间的距离或相似度, 再根据点之间的距离定义类间距离. 常用的点间距离有欧氏距离(Euclidean distance)、平方欧氏距离(squared Euclidean distance)、Chebychev距离、Minkovski距离、绝对距离(Block或absolute distance); 而相似度常用夹角余弦(cosine), Pearson相关系数等等, 其中夹角余弦和相关系数称为相似系数, 它们的值越大, 则说明距离越近. 而常用的类间距离定义包括最短距离法、最长距离法、重心法、类平均法、离差平方和法、中间距离法、可变平均法等等.

假定要确定p维点(向量)$(x_1, ..., x_p)$ 和$(y_1, ..., y_p)$之间的距离. 先介绍少数常用的点间距离公式.

距离或亲近度	公式
欧氏距离	$\sqrt{\sum_i (x_i - y_i)^2}$
平方欧氏距离	$\sum_i (x_i - y_i)^2$
绝对距离	$\sum_i \lvert x_i - y_i \rvert$
Chebychev距离	$\max_i \lvert x_i - y_i \rvert$
Minkovski距离	$\{\sum_i (x_i - y_i)^q\}^{1/q}$
夹角余弦	$\cos\theta_{xy} = \sum_i x_i y_i / \sqrt{\sum_i x_i^2 \sum_i y_i^2}$
Pearson相关系数	$r_{xy} = \sum_i (x_i - \overline{x})(y_i - \overline{y}) / \sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}$

假定要确定类G_p和类G_q之间的距离D_{pq}. 用$d(x_i, x_j)$表示在属于G_p的点x_i和属于G_q的点x_j之间的距离, 那么下面就是一些类间距离的定义方法.

- 最短距离法: $D_{pq} = \min d(x_i, x_j)$
- 最长距离法: $D_{pq} = \max d(x_i, x_j)$
- 重心法: $D_{pq} = d(\bar{x}_p, \bar{x}_q)$
- 类平均法: $D_{pq} = \frac{1}{n_1 n_2} \sum_{x_i \in G_p} \sum_{x_j \in G_q} d(x_i, x_j)$
- 离差平方和法: $D_{pq} = D_{1+2} - D_1 - D_2$, 这里定义$D_1 = \sum_{x_i \in G_p} (x_i - \bar{x}_p)^T (x_i - \bar{x}_p)$, $D_2 = \sum_{x_j \in G_q} (x_j - \bar{x}_q)^T (x_j - \bar{x}_q)$, $D_{1+2} = \sum_{x_k \in G_p \cup G_q} (x_k - \bar{x})^T (x_k - \bar{x})$.

k均值聚类: 先决定选择把观测值分成多少类(假定k类), 然后以(有些任意的)k个点为"种子", 按照到它们的距离远近把所有点分成k类, 再以这k类的均值(重心)为新的"种子"再重新分类, 如此下去, 直到收敛或者达到预定的迭代目标, 得到最终的k类.

分层聚类: 从每个点都看成一类开始进行两两合并, 每次合并距离最近的两类直到只有一类为止. 最后再根据需要, 按照结果(比如树状图), 得到分类.

3. 典型相关分析

对于两组变量$X = (X_1, ..., X_p)^T$和$Y = (Y_1, ..., Y_p)^T$, 寻找和它们有关两个系数向量$a = (a_1, ..., a_p)^T$和$b = (b_1, ..., b_q)^T$ 使得新的称为典型变量的综合变量

$$V_1 = a^T X = a_1 X_1 + a_2 X_2 + \cdots + a_p X_p$$
$$W_1 = b^T Y = b_1 Y_1 + b_2 Y_2 + \cdots + b_q Y_q$$

有尽可能大的相关关系. 令

$$Z = \begin{pmatrix} X \\ Y \end{pmatrix}, \quad Cov(Z) \equiv \Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}$$

从数学上来说, 典型相关分析的问题实际上是在约束条件$Var(V) = Var(W) = 1$下寻求a和b使得相关系数(这时等于协方差) $\rho_{VW} = Cov(V, W) = a^T \Sigma_{VW} b$最大. 这涉及解两个有同样(数目均为$k = \min(p, q)$, 而且取值于0和1之间的)非零特征值($\lambda^2$)的特征值问题:

$$Aa = \lambda^2 a, Bb = \lambda^2 b,$$

这里

$$A \equiv \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}, \ B \equiv \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}.$$

记A和B的非零特征根和特征向量为

$$\lambda_1^2 \geq \lambda_2^2 \geq \cdots \lambda_k^2 > 0 \text{ 和 } a_{(1)}, ..., a_{(k)}, b_{(1)}, ..., b_{(k)}$$

可得到k对线性组合$V_i = a_{(i)}^T X$, $W_i = b_{(i)}^T Y$, $i = 1, ..., k$. 每一对变量(V_i, W_i)称为典型变量. 比如, 最大特征值的平方根为V_1和W_1之间的相关系数$\rho_{V_1 W_1} =$

$Corr(V_1, W_1) = \lambda_i$, 称为第一典型相关系数. 一般只取前几个影响大的典型变量和典型相关系数来分析.

典型变量的性质为:

(1) X和Y本身的一切典型变量都不相关, 即X的典型变量$V_1, V2, ...$等等都不相关, 而Y的典型变量$W_1, W_2, ...$等等也都不相关.

(2) X和Y的同一对典型变量V_i和W_i之间的相关系数为λ_i, 不同对的V_i和$W_j (i \neq j)$之间不相关.

当然在实际例子中一般并不知道Σ. 因此在只有样本数据的情况下, 只要把Σ用样本协差阵或样本相关阵代替就行了. 但是这时的特征根可能不在0和1的范围, 因此会出现软件输出中的特征根(有可能远远大于1)不等于相关系数的平方的情况, 一般, 各种软件会给出调整后的相关系数.

4. **对应分析**

从原理上来说, 对应分析把一对列变量和一对行变量同时反映到同一张图上, 各自有其相应的坐标轴(因子轴). 下面从数学的角度解释对应分析. 假定数据矩阵为$n \times m$矩阵$A = \{a_{ij}\}$, 或者说行变量有n个水平(行变量), 列变量有m个水平. 为了要把行变量和列变量关联起来, 将用两个向量来代表行和列变量, 分别称为**行记分(row score)**和**列记分(column score)**. 令行记分为一个n维向量$x = \{x_i\}$, 而列记分为一个m维向量$y = \{y_j\}$. 那么对于满足下面条件的三元组(r, x, y)称为对应分析问题$C_0(A)$的解:

$$rx_i = \sum_{j=1}^{m} \frac{a_{ij} y_j}{a_{i\cdot}}, \quad i = 1, ..., n,$$

$$ry_j = \sum_{i=1}^{n} \frac{a_{ij} x_i}{a_{\cdot j}}, \quad j = 1, ..., m,$$

式中, $a_{i\cdot} = \sum_j a_{ij}, a_{\cdot j} = \sum_i a_{ij}$是各行及各列的元素和. 该式的意义为: 行记分(row score) x_i与列记分(column score) y_j的加权均值成比例, 而列记分y_j则与行记分x_i的加权均值成比例. 数值r为行列记分的相关(在典型相关的意义上). 对该问题解的数学推导也很简单, 对A, x, y做某种变换: 记

$$R \equiv \text{diag}(a_{i\cdot}), \ C \equiv \text{diag}(a_{\cdot j}), \ R^{1/2} \equiv \text{diag}(a_{i\cdot}^{1/2})$$

则上面式子有下面矩阵(向量)形式

$$rx = R^{-1} A y; \ ry = C^{-1} A^T x,$$

这里diag(w_i)代表由向量$\{w_i\}$作为对角线元素的对角线矩阵. 而$a_{i\cdot}$与$a_{\cdot j}$分别代表A的行总和及列总和所形成的向量. 不难验证这两个式子等价于

$$rR^{1/2} x = (R^{-1/2} A C^{-1/2}) C^{1/2} y;$$
$$rC^{1/2} y = (C^{-1/2} A^T R^{-1/2}) R^{1/2} x = (R^{-1/2} A C^{-1/2})^T R^{1/2} x.$$

由此, x为该方程一个解的条件是下面两组特征值问题有解[1]:

$$r^2(R^{1/2}x) = (R^{-1/2}AC^{-1/2})(R^{-1/2}AC^{-1/2})^T(R^{1/2}x);$$
$$r^2(C^{1/2}y) = (R^{-1/2}AC^{-1/2})^T(R^{-1/2}AC^{-1/2})(C^{1/2}y).$$

令

$$Z \equiv R^{-1/2}AC^{-1/2},\ v \equiv R^{1/2}x,\ u \equiv C^{1/2}y,$$

前面的特征值问题可以写成

$$r^2u = Z^TZu \text{ 及 } r^2v = ZZ^Tv.$$

这是两个特征值问题. 它们有同样的非零特征值. 如U是Z^TZ的特征向量, 则ZU是ZZ^T的特征向量. 根据线性代数, 显然r^2为对应于行和列的两个特征值问题的共同最大特征值的解(这里取最大的两个, 最多不超过$\min(m,n)$个). 此后的分析就和主成分分析等类似了, 也有载荷图, 但由于是两个有同样非零特征值的特征值问题, 就会有两个载荷图, 这两个载荷图重叠展示则产生对应分析的图(如图8.9).

8.5.2　R语句的说明

由于多数例题的R代码已经在课文中展示, 这里不重复. 仅给出一些补充.

1.　主成分分析

在前面课文中, 我们做主成分分析完全按照数学公式做的, 没有用现成的函数, 下面语句是对例8.1用现成函数princomp()的代码:

```
w=read.table("who.txt",sep=",",header=T)
y=princomp(w,cor=T);
y$sdev #特征值的平方根
y$load #单位特征向量等
y$scores #因子得分
screeplot(y) #画scree图
sweep(y$loa[,1:10],2,y$sde,"*")#单位特征向量乘相应特征值平方根得相关系数
```

2.　典型相关分析

在前面课文中, 我们做典型相关分析时用的是程序包CCA中的函数, 下面语句是对例8.4用另一个函数cancor()的代码:

```
w=read.table("tv.txt",header=T)
```

[1]最大特征值为1是平凡解, 两组的非零特征值相同!

```
X=w[,1:3];Y=w[,4:6]
a=cancor(X,Y)
a$xcoef #给出了第一组变量的三个典型变量的系数
a$ycoef #给出了第二组变量的三个典型变量的系数
a$cor #给出了这三对变量的相关系数
```

8.6 习题

1. 重复例8.1(数据who.txt)的主成分分析和因子分析的计算. 说明它们的区别.

2. 重复例8.1(数据who.txt)的主成分分析和因子分析的计算, 但试用不同选项. 看结果和第1题是否有区别.

3. 利用例7.1 (数据: bschool.txt)的美国60个著名商学院的数据, 包括的变量有GMAT分数、学费、进入MBA前后的工资等等, 其中有4个定量变量. 试图对这4个变量用主成分分析进行降维. 得到结果后, 再对该数据做因子分析. 比较这两个结果, 得出你的结论.

4. 对本书所附数据student.txt进行主成分分析和因子分析的计算, 解释结果, 说明它们的区别. 注: 那里的数据为100个学生的数学、物理、化学、语文、历史、英语的成绩.

5. 重复对例8.2的分层聚类, 只不过去掉一两个变量, 看聚类过程和结果是如何变化的.

6. 对例8.1的数据(who.txt)进行分层聚类(R型聚类), 分成几类合适? 并试图解释聚类结果.

7. 对例8.3做快速聚类或两步聚类, 比较结果.

8. 把例8.4数据(tv.txt)的各个变量重新组合分成两组(比如把hed、arti和man分成一组, 而led、net和com分在另一组)进行典型相关分析, 看典型相关系数如何变化, 并对照例8.4来解释结果.

9. 举出实际中可能应用典型相关分析的例子.

10. 对应分析和因子分析有什么不同?

11. R程序包MASS有一个caith数据(也是头发和眼睛颜色的数据)对其做对应分析, 和用例7.8做的对应分析结果做比较, 并解释输出.

第九章　随时间变化的对象: 时间序列分析

能不能用一个商场前12个月的销售情况来预测其下个月的销售额? 能不能用过去5年的月度物价指数来预测明年的物价指数? 这些问题所研究的对象都和时间有关系, 也就是本章要介绍的**时间序列(time series)**. 时间序列模型也是某种回归模型, 但方法和以前介绍的回归有很大区别, 它是用同一个变量过去的观测值来预测其未来的观测值.

人们对统计数据往往可以根据其特点从两个方面来切入, 以简化分析过程. 一个是研究所**谓横截面(cross section)**数据, 也就是研究对大体上同时发生的或者和时间关系不大的不同对象的观测值组成的数据. 另一个就是时间序列, 也就是由对象在不同时间的观测值形成的数据. 前面讨论的模型多是和横截面数据有关. 这里所说的时间序列的时间间隔是固定的, 而且观测时间有一定的长度. 有些数据也有很多重复观测, 但观测时间间隔不一定一样, 而且重复次数较短, 而且可能会有多个变量, 称为纵向数据. 纵向数据的处理方法和这里要讲的时间序列完全不同, 本书将不予以讨论. 本书主要讨论一个变量的时间序列, 不去讨论同时处理多个时间序列的问题.

经典的回归分析的目的是建立因变量和自变量之间关系的模型, 并且可以用自变量来对因变量进行预测. 经典线性回归分析模型中的误差项通常假定是互相独立并且有同样分布. 而时间序列的观测值并不独立, 比如一个企业今天的收入和其昨天的收入就很相关. 时间序列的因变量为变量未来的可能值, 而用来预测的自变量中就包含该变量的一系列历史观测值. 下面看一个时间序列的数据例子. 希望能够从这个数据找出一些规律, 并且建立可以对未来进行预测的时间序列模型.

例9.1 税收数据(tax.txt) 这是某地从1995年1月到2005年7月的税收(单位: 万元). 该数据为按照时间顺序的按月记录, 共127个观测值. 从该数据中的众多的数目只能够看出一个大概, 即总的趋势是增长, 但有起伏. 利用点图则可以得到对该数据更加直观的印象. 图9.1就是由该数据得到的一个时间序列图. 从这个点图可以看出, 总的趋势是增长的, 但增长并不是单调上升的. 大体上看, 这种升降不是杂乱无章的, 和季节或月份的周期有关系. 当然, 除了增长的趋势和季节影响之外, 还有些无规律的随机因素的作用. 这个只有一种随着时间变化的变量的序列一般称为**纯粹时间序列(pure time series)**. 下面将通过该例子对纯粹时间序列进行介绍.

该图是用下面代码画的(包括数据输入和对数据定义起始时间和周期):

```
x=scan("tax.txt")
tax=ts(x, frequency = 12, start = c(1995, 1))
ts.plot(tax,ylab="Tax")
```

聪明的读者可能马上会问, 用一个变量本身的历史值来预测其未来值会准确吗? 这种想法是非常有道理的. 仅仅在孤立系统, 也就是说, 其他因素对感兴趣的变量没有影响或者影响可以抵消或者忽略时, 时间序列分析才有意义. 大家可以想

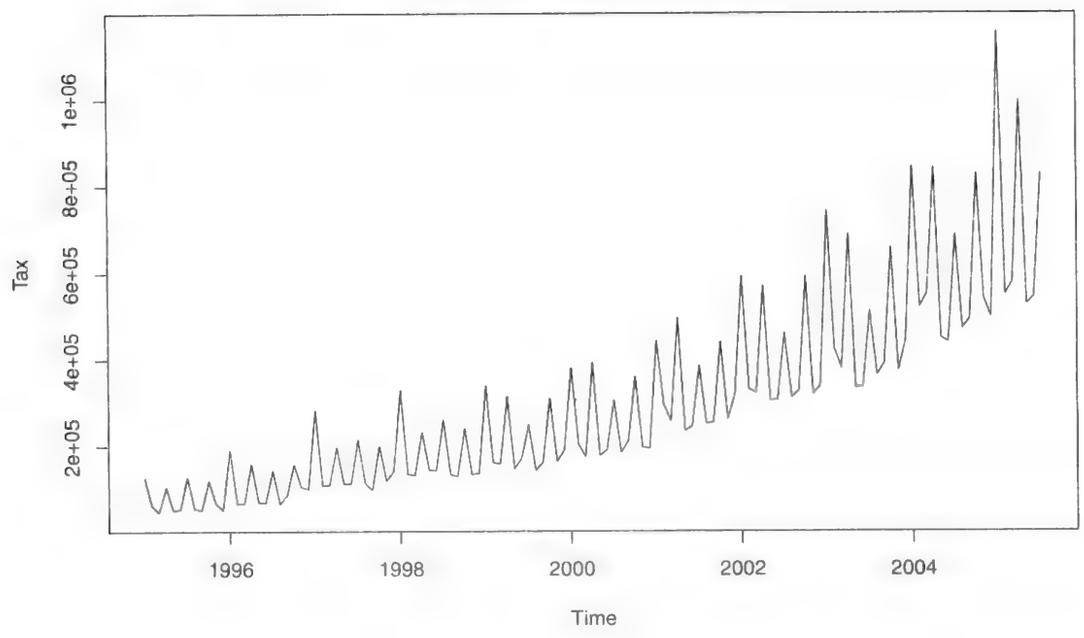

图 9.1　某地从1995年1月到2005年7月的税收数据图(单位: 万元).

一想, 什么类型的变量可能具有这种孤立的性质呢? 实际上, 纯粹孤立系统是不存在的, 所以, 时间序列仅仅是现实世界的一种近似. 任何统计模型都有其局限性, 都是对实际世界的不同程度的近似. 统计学家不应对任何统计模型的功效予以夸大, 更不能把一种模型或者理论当成信仰.

9.1　时间序列的组成部分

从图9.1可以看出, 该时间序列总体上是有一个上升趋势, 但又有些周期性波动, 但又不是非常整齐的. 一般来说, 比较简单的时间序列可以有三部分组成: **趋势(trend)**、**季节(seasonal)**成分和无法用趋势和季节模式解释的**随机干扰(disturbance)**[1]. 例9.1数据的税收就可以用这三个成分叠加而成的模型来描述. 一些时间序列还可能有**循环或波动(Cyclic, or fluctuations)**成分, 循环模式和有规律的季节模式不同, 周期长短不一定固定. 比如经济经济危机周期、金融危机周期等等. 一个时间序列可能有趋势、季节、循环这三个成分中的某些或全部再加上随机成分. 因此, 如果要想对一个时间序列本身进行较深入的研究, 把序列的这些成分分解出来, 或者把它们过滤掉则会有很大的帮助. 如果要进行预测, 则最好把模型中的与这些成分有关的参数估计出来. 对例9.1的时间序列通过计算机软件进行分解, 则可以轻而易举地得到该序列的趋势、季节和误差成分. 下面的图9.2的左图表示了去掉季节成分, 只有趋势和误差成分的序列的一条曲线. 图9.2中间图用两条曲线分别描绘了纯趋势成分和纯季节成分. 图9.2右图用两条

[1]随机干扰在模型中也称为误差.

曲线分别描绘了纯趋势成分和纯误差成分. 这些图直观地描述了对于带有几种成分的时间序列的分解. 该图是用下面代码画的:

```
a=stl(tax, "period") #进行分解
par(mfrow=c(1,3))
plot(tax-a$time.series[,1],ylab="",main="Without Seasonal")
ts.plot(a$time.series[,1:2],main="Trend and error")
ts.plot(a$time.series[,2:3],main="Trend and Seasonal")
```

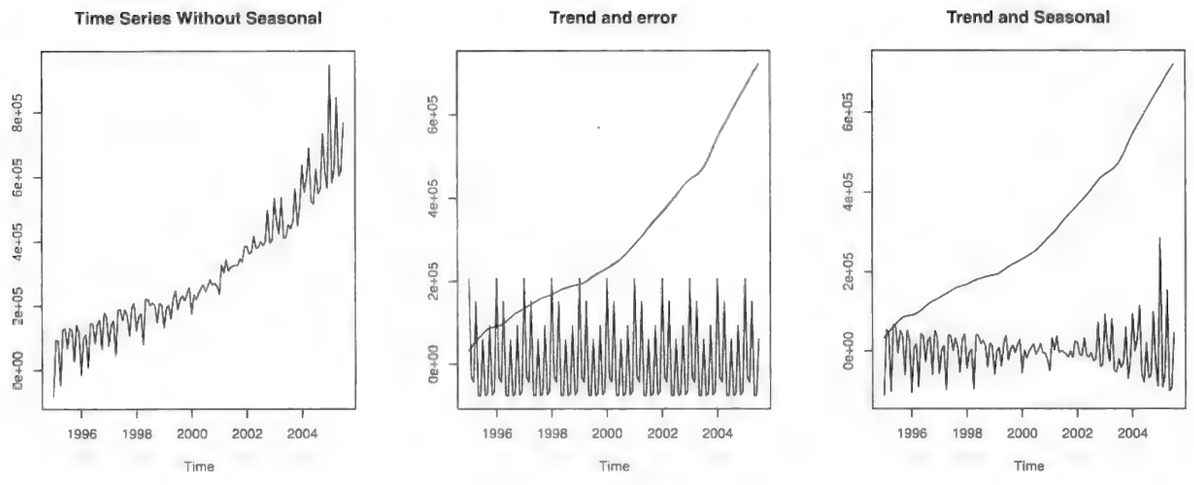

图 9.2　例9.1数据的分解图.

可以看到, 在定义时间序列时, 周期必须事先确定, 单位为年、月、日、周等等. 假定感兴趣的是零售商品的销售, 如果有特殊日子(比如像新年及圣诞节一类的在阳历中固定的节假日), 则必须予以方法上的调整, 而对于在阳历中不固定的节假日(如春节、中秋节等)就没有那么简单了.

9.2　指数平滑

如果人们不仅仅满足于分解现有的时间序列, 而且想要对未来进行预测, 就需要建立模型. 首先, 这里介绍比较简单的**指数平滑(exponential smoothing)**. 指数平滑只能用于纯粹时间序列的情况, 指数平滑的原理为: 当利用过去观测值的加权平均来预测未来的观测值时(这个过程称为平滑), 离得越近的观测值要给以更多的权. 而"指数"意味着: 按照已有观测值"陈旧"程度增加的方向, 在其上所加的权数按指数速度递减. 以简单的没有趋势和没有季节成分的纯粹时间序列为例, 指数平滑在数学上实际是一个几何级数. 这时, 如果用Y_t表示在t时间的平滑后的数据(或预测值), 而用$X_1, X_2, ..., X_t$表示原始的时间序列. 那么指数平滑最简单的模型为

$$Y_t = \alpha X_t + (1-\alpha)Y_{t-1}, \quad 0 < \alpha < 1.$$

这里当$t = 1$时, 会出现未知的Y_0, 它是需要设定的初始值, 通常设为X_1, 这时$Y_1 = X_1$. 该模型可等价地写成$Y_t = \alpha \sum_{k=0}^{t-1}(1-\alpha)^k X_{t-k}$. 这里的系数为几何

级数. 因此有人认为称之为"几何平滑"比使人不解的"指数平滑"似乎更有道理. 自然, 这种在简单情况下导出的公式(如上面的公式)无法应对具有各种成分的复杂情况. 本章后面将给出各种实用的指数平滑模型的公式. 根据数据, 可以得到这些模型参数的估计以及对未来的预测. 在和例9.1有关的指数平滑模型中, 需要估计12个季节指标和三个参数(包含前面公式权重中的α、和趋势有关的γ以及和季节指标有关的δ). 在简单的选项之后, 可以利用计算机软件通过指数平滑产生对2005年7月后一年的预测. 图9.3为用R软件绘出的原始的时间序列(实线)和预测的部分(虚线), 包括对2005年7月之后12个月的预测. 为对例9.1进行指数平滑以及预测并形成图9.3使用了下面代码:

```
b=HoltWinters(tax,beta=0);tax.p=predict(b,n.ahead=12)
ts.plot(tax,xlim=c(1995,2006.5));lines(tax.p,col=1,lty=3)
```

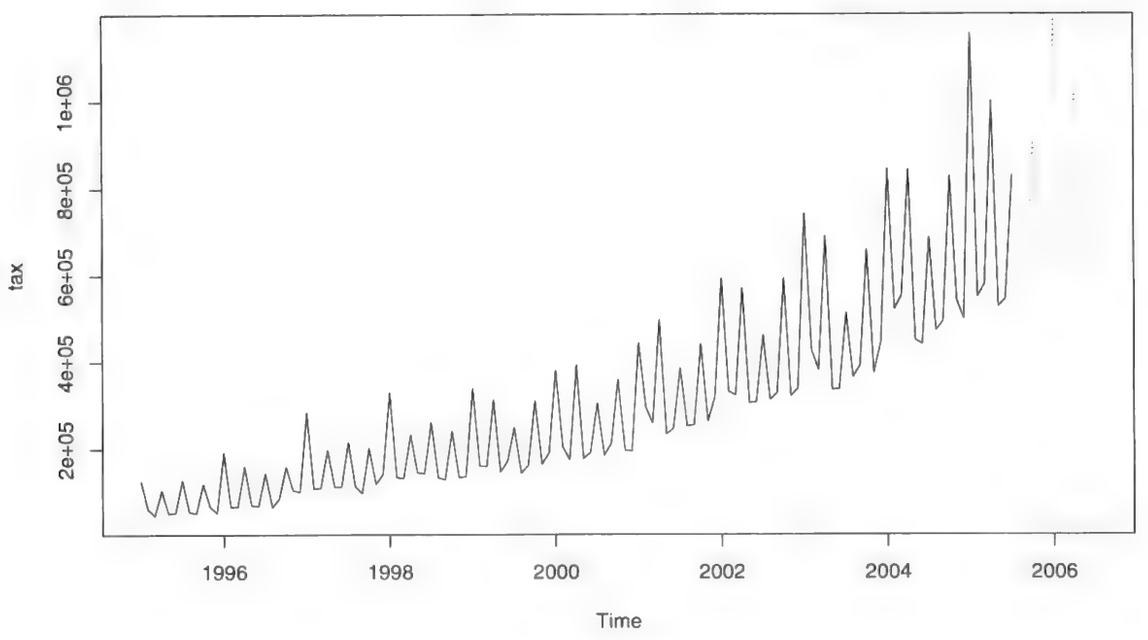

图 9.3 例9.1数据通过指数平滑做出12个月的预测.

如果要对比较复杂的纯粹时间序列进行细致的分析, 指数平滑并不总是满足要求的. 人们想出了数学上更加复杂的模型, 这就是下面要介绍的Box-Jenkins ARIMA模型.

9.3 Box-Jenkins 方法: ARIMA模型

9.3.1 ARIMA模型介绍

比指数平滑要更精细的模型是Box-Jenkins引入的ARIMA模型, 或称为**整合自回归移动平均模型(ARIMA** 为Autoregressive Integrated Moving Aver-

age的些关键字母的缩写). 前面说过, 一个时间序列可能会由季节、趋势和随机干扰三个部分组成. ARIMA的想法是, 如果能够把季节和趋势从数据中过滤掉, 而剩下的部分有某些数学上的特点, 使得可以对剩下的那部分建模, 即后面要介绍的ARMA模型, 那么, 在得到该模型之后, 再把周期和趋势整合进去(把字母I加到ARMA中)以得到结果, 这就是ARIMA模型名字的来历.

ARIMA模型的基础是**自回归和移动平均模型**或**ARMA(Autoregressive and Moving Average)** 模型. 它由两个特殊模型发展而成, 一个特例是**自回归模型**或**AR**(Autoregressive)模型. 如果时间序列用$X_1, X_2, ..., X_t$表示, 则一个纯粹的AR(p) 模型意味着变量的一个观测值由其以前的p个观测值的线性组合加上随机误差项a_t(该误差为独立不相关的)而得:

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + a_t.$$

这看上去像序列自己对自己回归一样, 所以称为自回归模型. 它牵涉到过去p个观测值(相关的观测值间隔最多为p个). ARMA模型的另一个特例为**移动平均模型**或**MA(Moving Average)** 模型. 一个纯粹的MA(q)模型意味着变量的一个观测值由目前的和先前的q个随机误差的线性的组合:

$$X_t = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}.$$

由于右边系数的和不为1(系数甚至不一定是正数), 因此有人觉得, 叫做"移动平均"不如叫做"移动线性组合"更确切, 虽然行家已经习惯于叫"平均"了, 但初学者还是因此可能和初等平滑方法中的什么"三点平均"之类的术语混淆. 显然, ARMA(p, q)模型应该是AR(p)模型和MA(q)模型的组合了:

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}.$$

显然ARMA($p, 0$)模型就是AR(p)模型, 而ARMA($0, q$)模型就是MA(q)模型. 一般的ARMA(p, q)模型有$p + q$个参数要估计, 看起来很繁琐, 但利用计算机软件则是常规运算, 并不复杂. 但是要想ARMA(p, q)模型有意义则要求它必须满足**平稳性(stationarity)**和**可逆性(invertibility)**的条件, 这意味着序列均值不随着时间增加或减少、序列的方差不随时间变化、序列本身相关的模式不改变等等许多数学条件. 一个实际的时间序列是否满足这些条件是无法在数学上验证的. 一个模型是否可用, 可以通过下面要介绍的时间序列的自相关函数图和偏相关函数图中大体识别出来. 一般人们所关注的有趋势和季节/循环成分的时间序列都不会满足这些要求的. 这时就需要对时间序列进行**差分(difference)**来消除这些使序列不平稳的成分, 而使其变成平稳的时间序列, 并估计ARMA模型的参数, 估计之后再转变该模型, 使之适应于差分之前的序列(这个过程和差分相反. 所以称之为整合的(integrated)ARMA模型), 得到的最终模型于是称为ARIMA模型.

这里所说的差分是什么意思呢? 差分可以是每一个观测值减去其前面的一个观测值, 即

$$X_t - X_{t-1}$$

这样, 如果时间序列有一个斜率不变的趋势, 经过这样的差分之后, 该趋势就会被消除了. 当然差分也可以是每一个观测值减去其前面任意间隔的一个观测值, 比

如时间序列存在周期为s的季节成分, 那么相隔s的差分

$$X_t - X_{t-s}$$

就可以把这种以s为周期的季节成分消除. 对于复杂情况, 可能要进行多次差分, 才能够使得变换后的时间序列平稳和可逆. 当然, 也可能永远达不到这种要求.

思考一下:

1. ARIMA模型是通过差分, 把时间序列变成ARMA模型, 再用数学方法来得到它, 然后再整合(Integration)成ARIMA模型. 这主要因为, 人们在数学上对假定的ARMA模型还有些办法.
2. ARMA模型需要假定很多无法由数据来验证的数学条件, 因此, 只能希望这里的序列经过差分之后能够近似地满足这些条件. 但现实世界毕竟和理想的数学世界有差距. 完全符合现实世界的模型是不存在的.

9.3.2 ARMA模型的识别和估计

上面一小节, 引进了一些必要的术语和概念. 下面就如何识别模型进行说明. 要想拟合ARIMA模型, 必须先把它利用差分变成ARMA(p,q)模型, 并确定是否平稳, 然后确定参数p和q. 现在利用一个例子来说明如何识别一个AR(p)模型和参数p. 而MA(q)及ARMA(p,q)模型可用类似的方法来识别. 根据ARMA(p,q)模型的定义, 它的参数p,q的取值大小和**自相关函数(acf, autocorrelations function)**及**偏自相关函数(pacf, partial autocorrelations function)**有关. 自相关函数描述观测值和前面的观测值的相关系数, 而偏自相关函数为在给定中间观测值的条件下观测值和前面某间隔的观测值的相关系数. 举例来说, acf图上在整数横坐标0, 1, 2,...上有许多条, 在第i坐标上的条的高度等于X_t和X_{t-i}的相关系数, 而在pacf图上第i个条的高度等于X_t与X_{t-i}在其中间的值给定的条件下的相关系数. 在这两个图上如果横坐标有0(有时没有), 那么上面条的高度应该是1(X_t和X_t自己的相关系数). 这里当然不打算讨论这两个概念的细节. 引进这两个概念主要是为了能够了解如何通过研究关于这两个函数的acf和pacf图来识别模型. 为了直观地理解上面的概念, 下面利用一个例子(例9.2)来描述.

例9.2 数据(ar2.txt) 该数据是为了说明如何对一个时间序列数据进行AR模型识别. 原始时间序列由图9.4(上图)描述. 该序列的acf和pacf图显示在图9.4的下面左右两图. 图9.4是由下面代码(包括读入数据)实现的:

```
x=scan("ar2.txt")
layout(matrix(c(1,1,2,3),nr=2,byrow=T))
ts.plot(x);acf(x);pacf(x)
```

图9.4下面左图的acf条形图是衰减的指数型的波动, 这种图形称为拖尾. 而右边的pacf条形图是在第二个条($p=2$)之后就很小, 而且没有什么模式, 这种图形

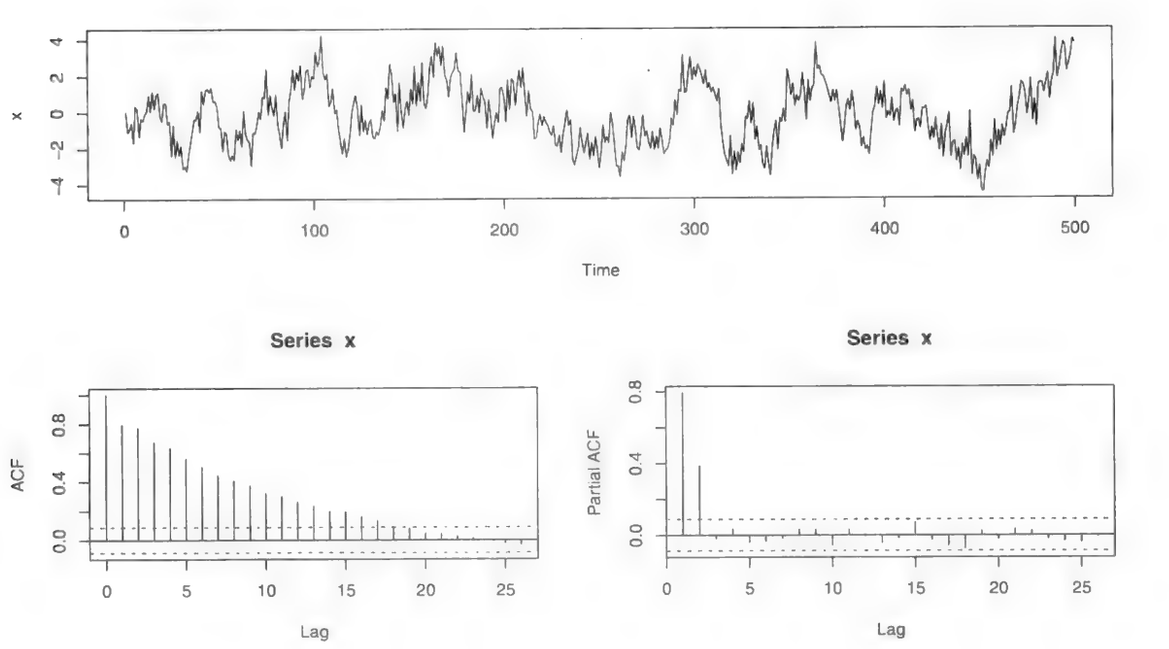

图 9.4　例9.2数据的时间序列(上图)及其acf(下左)和pacf(下右)图.

称为在$p=2$后截尾. 这说明该数据近似满足平稳的AR(2)模型. 注意, 所谓拖尾图形模式可能是以正负相间的正弦形式衰减, 也可能是以指数率衰减. 类似地, 如果acf图形是在第$q=k$个条后截尾, 而pacf图形为拖尾, 则数据可能适合MA(q)模型. 如果两个图形都拖尾则可能满足ARMA(p,q)模型. 具体的近似判别法总结在下面:

如何用pacf及acf图的拖尾和截尾判断ARMA模型

模型	AR(p)	MA(q)	ARMA(p,q)
pacf图形	在第p条后截尾	拖尾	头p条无规律, 其后拖尾
acf图形	拖尾	在第q个条后截尾	头q个条无规律, 其后拖尾

如果acf和pacf的图中均没有截尾, 而且至少有一个图没有显示以指数形式或正弦形式衰减, 那么说明该序列不是平稳序列, 必须进行差分变换来得到一个可以估计参数的满足ARMA(p,q)模型的序列. 而如果一个时间序列的acf和pacf图没有任何模式, 而且数值很小, 那么这个序列可能就是一些互相独立的无关的随机变量, 一个拟合良好的时间序列模型的残差就应该有这样的acf和pacf图.

图9.5为模拟的AR(2), $MA(2)$和ARMA(2,2)三个序列所对应的acf和pacf图. 注意, 图中有些条是从0开始的(不算在p或q内). 这几个图是用R软件模拟出来的. 可以看出, 上面表中的准则不那么准确, 按照上表来判断头两行为AR(2)和MA(2)没有问题, 但最后两图按照上表实在不好判断其p,q是多少. 这也说明这个判别的粗糙性. 特别是对ARMA模型, "头几个条无规律"这句话不那么清楚, 因此也不易掌握好, 很难判别准确. 模拟这几个序列及画相应的acf图及pacf图的R代码如下:

```
set.seed(1010)
x1=arima.sim(list(c(2,0,0),ar=c(0.3,-0.6)),n = 200)
x2=arima.sim(list(c(0,0,2),ma=c(-0.3,-0.4)),n = 200)
x3=arima.sim(list(c(2,0,2),ar=c(.3,.6),ma=c(.5,.2)),n=200)
par(mfrow=c(3,2))
acf(x1,main="Acf of AR(2) Series")
pacf(x1,main="Pacf of AR(2) Series")
acf(x2,main="Acf of MA(2) Series")
pacf(x2,main="Pacf of MA(2) Series")
acf(x3,main="Acf of ARMA(2,2) Series")
pacf(x3,main="Pacf of ARMA(2,2) Series")
```

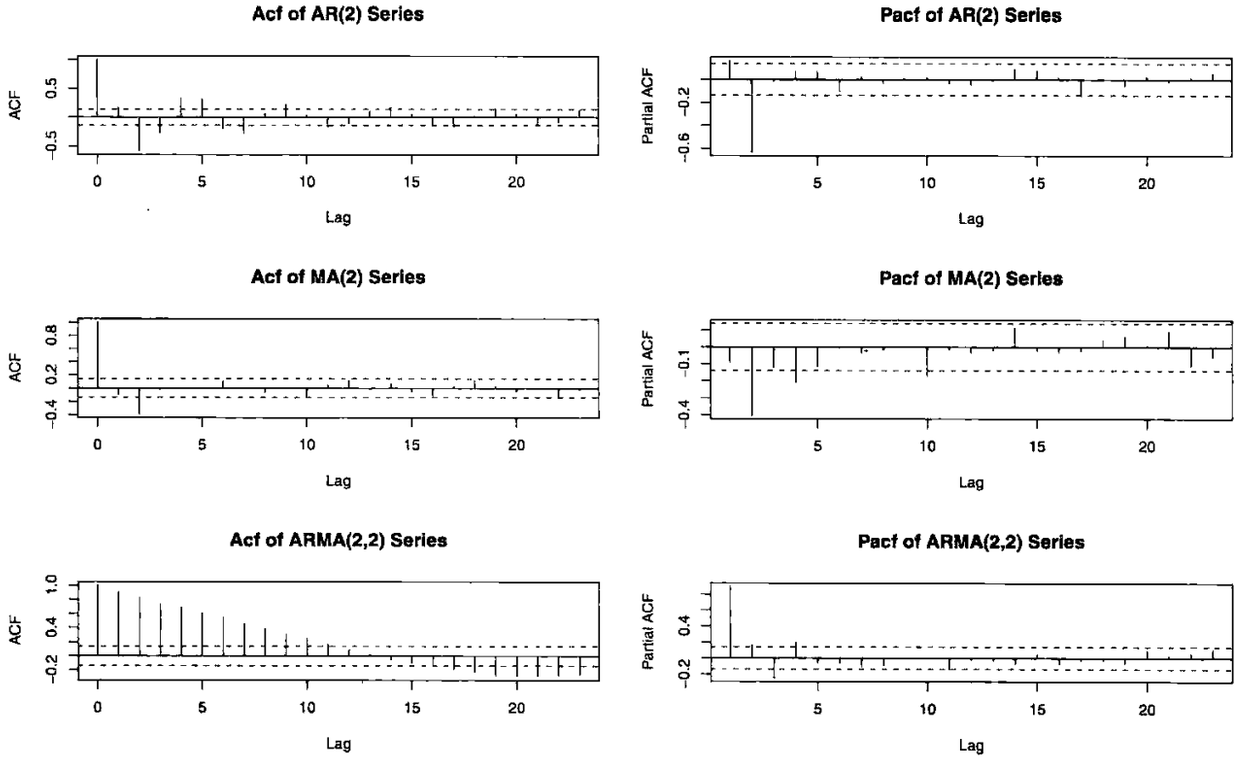

图 9.5　模拟的AR(2), MA(2)和ARMA(2, 2)序列所对应的acf和pacf图.

对于例9.2数据, 根据图9.4中acf和pacf图的形态, 不用进行任何差分就可以直接用AR(2)模型拟合, 下面R代码就包括了读入数据, 估计模型参数及对未来50个观测进行预测并画图(这里没有显示图形, 因为对平稳序列预测没有意义).

```
x=scan("ar2.txt");(d=arima(x,c(2,0,0),include.mean=F))
pr=predict(d,50);ts.plot(x);lines(pr$pred,lty=2)
```

输出为:

```
Coefficients:
```

```
          ar1      ar2
      0.4784   0.4064
s.e.  0.0408   0.0410
sigma^2 estimated as 0.8644:log likelihood=-673.76,aic=1353.51
```

得到的AR(2)参数估计为$\hat{\phi}_1 = 0.4784, \hat{\phi}_2 = 0.4064$, 也就是说拟合出来的该AR(2)模型为

$$X_t = 0.4784X_{t-1} + 0.4064X_{t-2} + a_t$$

其实, 对于平稳序列进行预测没有多大意义, 因为它们的均值是不变的. 下面再看剩下的残差序列是否还有什么模式. 这还可以由残差的acf和pacf条形图来判断. 这两个图分别在图9.6的左图和中图. 可以看出, 它们没有什么模式(注意acf图的第一个条是在0点), 这说明拟合比较成功. 图9.6右图为残差序列图, 从中看不出任何模式. 说明残差序列看来是(满足要求的)独立和随机的.

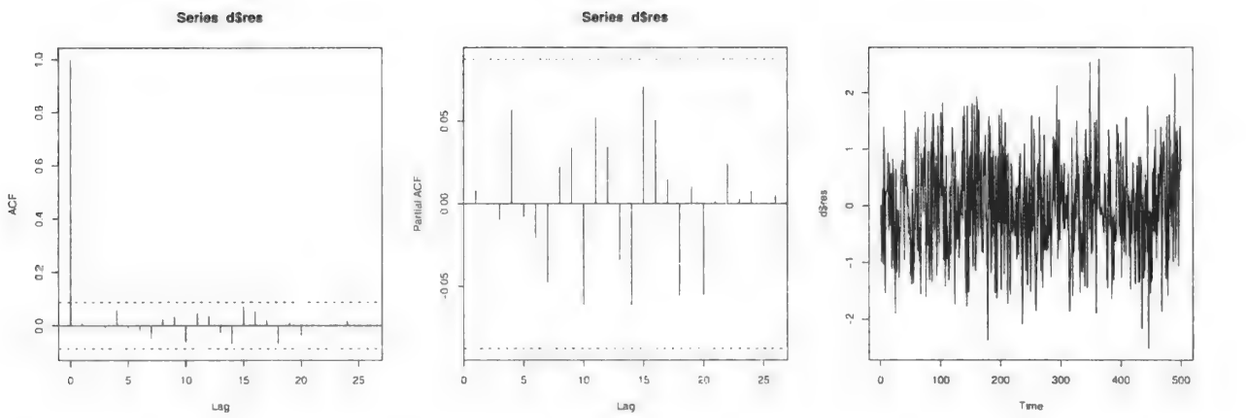

图 9.6　例9.2数据拟合AR(2)的残差序列的pacf(左)和acf(中)图及残差序列图(右).

思考一下:

1. 在拟合ARMA模型时也有许多不同的方法, 因此得到的估计结果也不尽相同. 在R中就有若干方法的选项, 还有包括不包括截距(或均值)的选项等等.
2. 在实际世界中, 很少会遇见平稳序列, 也就是方差和均值都不变的序列, 即使有, 也没有任何预测的必要, 因为平稳意味着均值不变, 还有什么可预测的呢?

9.3.3　用ARIMA模型拟合

在对含有季节和趋势/循环等成分的时间序列进行ARIMA模型的拟合研究和预测时, 就不像对纯粹的满足可解条件(平稳性和可逆性)的ARMA模型那么简单了. 一般的ARIMA模型有多个参数, 没有季节成分的可以记为ARIMA(p,d,q), 如果没有必要利用差分来消除趋势或循环成分时, 差分阶数$d = 0$, 模型为ARIMA$(p,0,q)$, 即ARMA(p,q). 在有已知的固定周期s时, 模型多了4个参

数, 可记为ARIMA$(p,d,q)(P,D,Q)^s$. 这里增加的除了周期s已知之外, 还有描述季节本身的ARIMA(P,D,Q)的模型识别问题. 因此, 实际建模要复杂得多. 需要经过反复比较.

先前对例9.1(数据tax.txt)进行了分解, 并且用指数平滑做了预测. 知道其中有季节和趋势成分. 下面试图对其进行ARIMA模型拟合. 先对该序列做acf和pacf条形图. 其中acf图(见图9.7)显然不是拖尾(不是以指数速率递减), 这说明需要进行差分. 关于参数的选择, 不要选得过大. 每次拟合之后要检查残差的acf和pacf图, 看是否为无关随机序列. 人们可能要经过多次对比, 才能把ARIMA模型的各个参数识别出来. 对于例9.1数据, 我们最后选中了ARIMA$(0,1,1)(1,2,1)^{12}$模型来拟合. 拟合的结果和对2005年7月之后12个月的预测在图9.8中.

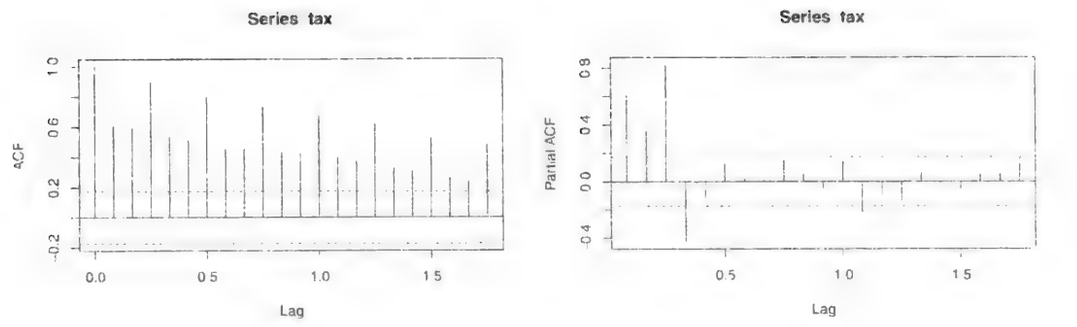

图 9.7 例9.1的时间序列的acf和pacf图.

绘制图9.7及拟合ARIMA$(0,1,1)(1,2,1)^{12}$模型的程序为(包括读入数据):

```
x=scan("tax.txt")
tax=ts(x, frequency = 12, start = c(1995, 1))
par(mfrow=c(1,2));acf(tax);pacf(tax)
(a=arima(tax,c(0,1,1),c(1,2,1)))
```

下面是给出了MA模型的一个系数、一个季节AR模型的系数及一个季节MA模型系数的估计的输出:

```
Series: tax
ARIMA(0,1,1)(1,2,1)[12]

Coefficients:
          ma1     sar1     sma1
      -0.8204  -0.5030  -0.5676
s.e.   0.0630   0.1233   0.1333

sigma^2 estimated as 1.213e+09:  log likelihood=-1223
AIC=2454.01   AICc=2454.42   BIC=2464.51
```

对以后12个月进行预测及产生图9.8的代码为

```
library(forecast)
fit <- Arima(tax,c(0,1,1),c(1,2,1))
plot(forecast(fit,h=12))
```

注意, 这里用了程序包forecast[1], 而且用Arima()函数重新做了拟合, 拟合结果当然和上面的一样, 用这个程序包主要是为了画图方便, 图中绘出了估计的置信带.

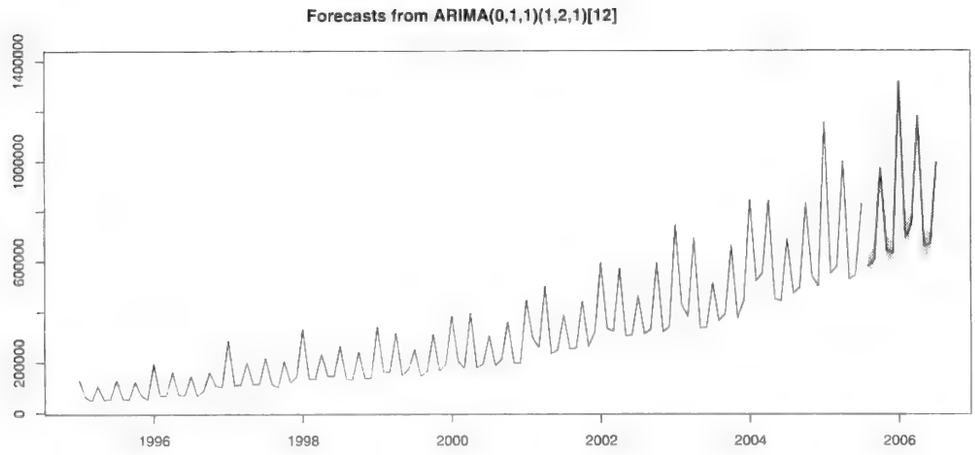

图 9.8　例9.1的原始序列和由模型得到的拟合值及对未来12个月的预测图.

为了核对, 当然要画出残差的acf和pacf的条形图来看是否还有什么非随机的因素存在. 图9.9为这两个图, 看来模型的选择还是适当的. 代码为`par(mfrow=c(1,2));acf(fit$res);pacf(fit$res)`.

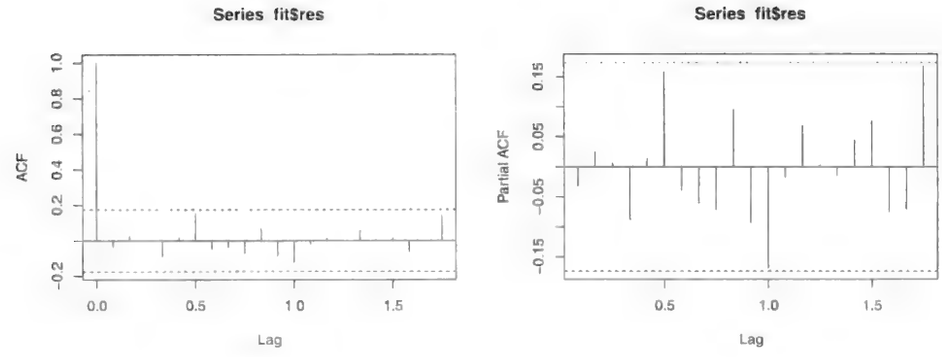

图 9.9　例9.1数据拟合ARIMA$(0,1,1)(1,2,1)^{12}$模型后残差序列的acf和pacf条形图.

[1]Rob J Hyndman with contributions from Slava Razbash and Drew Schmidt (2012). forecast: Forecasting functions for time series and linear models. R package version 3.25. http://CRAN.R-project.org/package=forecast.

在对模型的检验中, 有一个Ljung-Box检验, 其零假设为残差的各阶相关(下面式中写的最大k阶)等于零, 即

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_k = 0,$$

这里ρ_i表示i间隔的acf. 图9.10表示了对于例9.1拟合ARIMA$(0,1,1)(1,2,1)^{12}$模型后残差的Ljung-Box检验的p值. 看来都不显著. 也就是说, 至少在$k = 100$之前, 没有证据表明, 残差还有各阶自相关性. 绘制该图的R代码如下:

```
a=arima(tax,c(0,1,1),c(1,2,1)) #重复前面的拟合
B=NULL;for( i in 1:100)
B=c(B,Box.test(a$res,lag=i,type="Ljung-Box")$p.value)
plot(B,main="Ljung-Box tests", ylab="p-value",
xlab="lag",pch=16,ylim=c(0,1));abline(h=.05,lty=2)
```

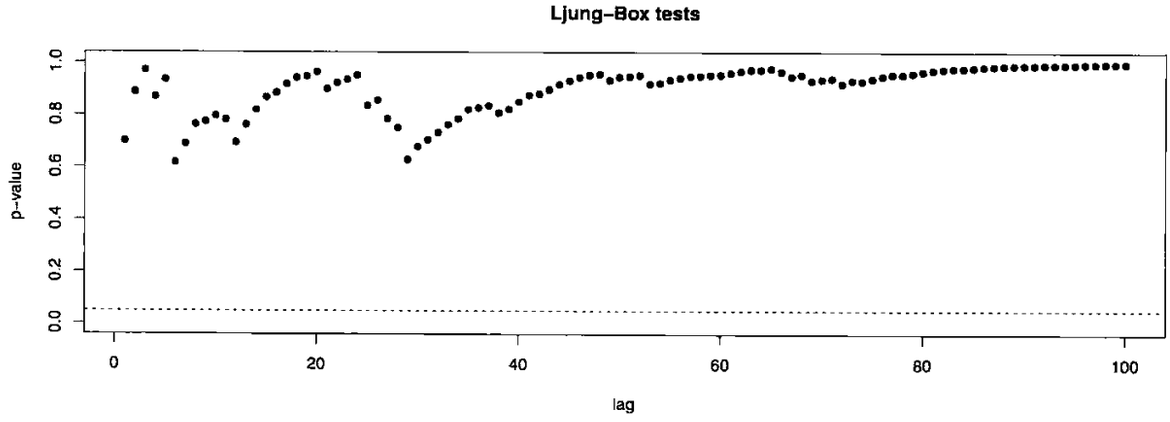

图 9.10 例9.1数据拟合ARIMA$(0,1,1)(1,2,1)^{12}$模型后残差序列的Ljung-Box检验的p值.

值得指出的是, 在各种软件中都有自动选择ARMA模型的程序或函数, 这些程序选择所使用的准则不尽相同, 选择的默认范围也不同, 因而结果也有所差异. 下面就用R程序包forecast所包含的自动拟合函数auto.arima()为例来介绍. 对我们的数据输入下面代码:

```
library(forecast)
(a1=auto.arima(tax))
```

马上得到拟合结果:

```
Series: tax
ARIMA(3,1,1)(0,1,0)[12]

Coefficients:
          ar1      ar2     ar3      ma1
      -0.0551  -0.0099  0.3470  -0.8881
```

```
s.e.    0.1194    0.1158  0.1098    0.0698

sigma^2 estimated as 1.329e+09:  log likelihood=-1359.98
AIC=2729.97    AICc=2730.52    BIC=2743.65
```

在其默认的准则和计算范围内, 它选择了ARIMA$(3,1,1)(0,2,0)^{12}$的模型, 显然和我们手工选的不同. 它的一些准则(这里指输出的AIC, BIC等等. 它们的值越小越说明模型在这些准则下越"好", 本书不予以介绍.)不如前面手工选择的"好". 也可以画出对于残差的Ljung-Box检验的p值图及acf和pacf图(图9.11), 似乎都不及前一个模型(比如一些p值较小, 一些pacf线过长等等). 但如果没有前面的模型, 谁也没有理由来否定第二个模型. 实际上, 所有的模型都是近似, 有些模型之间的好坏容易比较, 而另一些很难比较, 这也是同一个现象可能会有多个模型来说明的情况. 时间序列数学的假定太多, 比如线性形式、正态性等等, 而这些假定不可能被证明, 也不易判断.

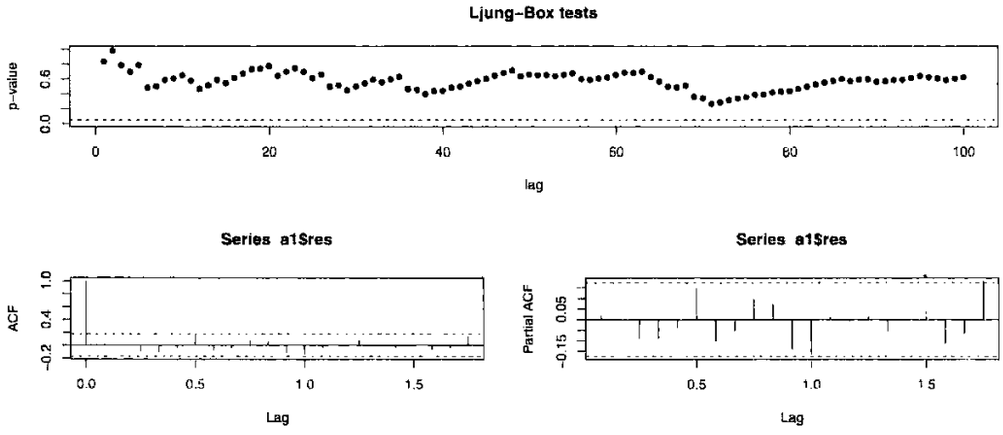

图 9.11　例9.1数据拟合自动选择的ARIMA$(3,1,1)(0,2,0)^{12}$模型后残差序列的Ljung-Box检验的p值.

9.4　小结

由于R的所有有关的语句都在前面相关地方写明了, 因此这里不再重复.

9.4.1　本章的概括和公式

本章引进了时间序列的概念, 并且介绍了用指数平滑和ARIMA模型来解时间序列的建模和预测问题. 这两个模型的主要公式如下, 仅供有兴趣的读者参阅.

1.　指数平滑模型

这些模型中有$\alpha,\gamma,\delta,\phi$为待估计参数, $\gamma=0$意味着斜率为常数(趋势无变化), 而$\delta=0$意味着没有季节成分, ϕ 和减幅趋势有关, p为季节周期; 对于时间序

列X_t来说, 趋势、光滑后的序列、季节因子和预测的序列分别用T_t, S_t, I_t 和$\hat{X}_t$表示; 另外, e_t为残差$X_t - \hat{X}_t$.

(1) **线性趋势可加季节模型(Linear trend, additive seasonality model)**为

$$X_t = b_0 + b_1 t + I_t + \epsilon_t$$
$$T_t = T_{t-1} + \alpha\gamma e_t,\ S_t = S_{t-1} + T_{t-1} + \alpha e_t,\ I_t = I_{t-p} + \delta(1-\alpha)e_t,$$
$$\hat{X}_t = S_t + T_t + I_{t-p+1}$$

(2) **线性趋势可乘季节模型(Linear trend, multiplicative seasonality model)**为

$$X_t = (b_0 + b_1 t)I_t + \epsilon_t$$
$$T_t = T_{t-1} + \alpha\gamma\frac{e_t}{I_{t-p}},\ S_t = S_{t-1} + T_{t-1} + \alpha\frac{e_t}{I_{t-p}},\ I_t = I_{t-p} + \delta(1-\alpha)\frac{e_t}{S_t},$$
$$\hat{X}_t = (S_t + T_t)I_{t-p+1}$$

(3) **指数趋势可加季节模型(Exponential trend, additive seasonality model)**为

$$X_t = b_0 b_1^t + I_t + \epsilon_t$$
$$T_t = T_{t-1} + \alpha\gamma\frac{e_t}{S_{t-1}},\ S_t = S_{t-1}T_{t-1} + \alpha e_t,\ I_t = I_{t-p} + \delta(1-\alpha)e_t,$$
$$\hat{X}_t = S_t T_t + I_{t-p+1}$$

(4) **指数趋势可乘季节模型(Exponential trend, multiplicative seasonality model)**为

$$X_t = (b_0 b_1^t)I_t + \epsilon_t$$
$$T_t = T_{t-1} + \alpha\gamma\frac{e_t}{I_{t-p}S_{t-1}},\ S_t = S_{t-1}T_{t-1} + \alpha\frac{e_t}{I_{t-p}},$$
$$I_t = I_{t-p} + \delta(1-\alpha)\frac{e_t}{S_t},\ \hat{X}_t = (S_t T_t)I_{t-p+1}$$

(5) **减幅趋势可加季节模型(Damped trend, additive seasonality model)**为

$$X_t = b_0 + \phi b_1 t + I_t + \epsilon_t$$
$$T_t = \phi T_{t-1} + \alpha(\alpha - \phi + 1)e_t,\ S_t = S_{t-1} + \phi T_{t-1} + \alpha(2-\alpha)e_t,$$
$$I_t = I_{t-p} + \delta[1 - \alpha(2-\alpha)]e_t,\ \hat{X}_t = S_t + \phi T_t + I_{t-p+1}$$

(6) **减幅趋势可乘季节模型(Damped trend, multiplicative seasonality model)**为

$$X_t = (b_0 + \phi b_1 t)I_t + \epsilon_t$$

$$T_t = \phi T_{t-1} + \alpha(\alpha - \phi + 1)\frac{e_t}{I_{t-p}},\ S_t = S_{t-1} + \phi T_{t-1} + \alpha(2-\alpha)\frac{e_t}{I_{t-p}},$$
$$I_t = I_{t-p} + \delta[1-\alpha(2-\alpha)]\frac{e_t}{S_t},\ \hat{X}_t = (S_t + \phi T_t)I_{t-p+1}$$

2. ARIMA模型

平稳时间序列X_t满足的条件: 对所有t, $E(X_t) = \mu$, 而且自协方差函数

$$\gamma_{ts} = Cov(X_t, X_s) = E(X_t - \mu)(X_s - \mu)$$

仅仅与$t-s$有关, 因此可以记$\gamma_k \equiv \gamma_{t,t+k} = Cov(X_t, X_{t+s})$. 对于平稳序列, 自相关函数(acf)定义为$Corr(X_t, X_{t+k}) = \gamma_k/\gamma_0$, 偏相关函数(pacf)定义为$Corr(X_t, X_{t+k}|X_{t+1}, ..., X_{t+k-1})$. 函数acf和pacf的点图可以用来帮助识别平稳过程的ARMA(p,q)模型. AR(p)和MA(q)模型是ARMA(p,q)模型的特例, 而ARMA(p,q)模型又是ARIMA(p,d,q)的特例(只有趋势, 没有季节), 而ARIMA(p,d,q)又是既有趋势又有季节成分的ARIMA$(p,d,q)(P,D,Q)^s$模型的特例. 为了便于描述公式, 定义算子

$$BX_t = X_{t-1},\ B^2X_t = X_{t-2}, ..., B^kX_t = X_{t-k}$$
$$(1-B^k)X_t = X_t - X_{t-k}$$
$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$
$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_p B^p$$

- **AR(p)模型:** $X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + a_t$, 或者用等价的算子符号, $\phi(B)X_t = a_t$.

- **MA(q)模型:** $X_t = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$, 或者用等价的算子符号, $\theta(B)a_t = X_t$.

- **ARMA(p,q)模型:** $X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$ 或者用等价的算子符号, $\phi(B)X_t = \theta(B)a_t$.

- **ARIMA$(p,d,q)(P,D,Q)^s$模型:**
$$\Phi_P(B^s)\phi_p(B)(1-B)^d(1-B^s)^D X_t = \theta_q(B)\Theta_Q(B^s)a_t,$$
这里Φ, Θ是类似于ARMA(p,q)模型中的算子ϕ, θ, 只不过是描述季节序列的罢了, 它们定义为
$$\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_p B^{ps}$$
$$\Theta_Q(B) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \cdots - \Theta_Q B^{Qs}$$

9.5 习题

1. 时间序列分析和一般的简单回归分析有什么不同?

2. 直观地说明时间序列中趋势和季节对序列的影响. 举出季节和趋势对你想象出来的任何时间序列的影响.

3. 举出实际中可能出现的时间序列.

4. 重复对书上例题的计算. 在选项上做一些变化, 试试自动选择模型的程序, 看结果有什么不同.

第十章　生存分析简介

很多人都可能会想过: "我到底能够活多少岁? "一些医生也会考虑: "到底这个新疗法能使得这类绝症患者多存活多久? "或者"还有什么别的因素和存活长短有关? " 之类的问题, 也可能会考虑, 一个病人在什么情况下的危险性最大等. 保险公司也要考虑各种人群的寿命, 以确保其人寿保险或医疗保险既具有竞争力又有利可图. 在工程上, 人们会考虑一个材料, 一个原件, 甚至一个设备的寿命是多少. 在经济活动中, 企业会关心员工在本公司能够工作多久而不跳槽, 什么因素会影响跳槽, 客户的忠诚状况能持续多久, 顾客的忠诚度会受什么因素影响. 在社会学中, 人们可能会考虑一个刑满释放人员, 能够持续多久不再犯罪, 或者在什么情况下最容易再次犯罪. 这些都属于统计中**生存分析(survival analysis)**的研究范围, 是研究一个事件在发生之前以什么概率持续多久的问题, 或者是什么时候, 以什么概率一个事件会发生. 本章主要介绍生存分析的一些基本知识, 并通过数据例子来介绍如何处理生存分析数据.

大家都明白, 对于某一特定个体"能够存活多久"这一类的问题, 任何负责任的人都不会作出确定的回答. 但是对于具有某些特性的一类人群, 则可以通过对数据的分析来近似地得到活过一定时间的概率. 如果关心不同治疗手段的效果, 还可以通过数据分析来比较这些方法, 看它们的有效性, 还能建立可以预测的量化的模型. 为此, 需要引进下面一些基本概念.

在生存分析中, 人们往往希望知道存活过时间t的概率, 这就是所谓的**生存函数(survival function)**, 记为$S(t)$, 显然它等于1减去生存时间不超过t的概率. 记$F(t)=P(T\le t)$为寿命不超过时间t的概率, 则$S(t)=1-F(t)$. 还要定义一个在t时刻处(附近), 对死亡发生的可能性进行度量的函数, 称为**危险函数(hazard function)**, 用$h(t)$表示, 它实际上是$-\ln S(t)$的关于t的导数, 代表了在活过了时间t的条件下, 在t时刻处死亡的(条件)概率密度函数, 或

$$h(t)=\lim_{\Delta t\to 0}\frac{P(t\le T\le t+\Delta t|T\ge t)}{\Delta t}=\lim_{\Delta t\to 0}\frac{F(t+\Delta t)-F(t)}{[1-F(t)]\Delta t}$$
$$=\frac{f(t)}{S(t)}=\frac{d}{dt}[-\ln S(t)].$$

累积危险函数为$H(t)=\int_0^t h(u)du$. 生存函数、危险函数、累积危险函数在数学上是等价的, 知道其中之一, 就可以推导出其他函数. 生存函数和危险函数可以形象地描述存活过某个时间的概率以及在各个时间段危险程度. 下面先看一个很有名的例子(Box and Cox, 1964).

例10.1 毒药数据(poison.txt) Box and Cox(1964)[1] 通过该数据引入了回归中的Box-Cox变换. 该数据是一个生存分析数据. 这个数据源于一个两个因素, 3×4水平的动物实验, 2个因素(自变量)为毒药(Poison, 3个水平), 处

[1]Box, G. E. P., and Cox, D. R. (1964). An analysis of transformations (with Discussion). *J. R. Statist. Soc.* B, 26, 211-252.

理(Treatment, 4个水平), 因变量为动物的生存时间(Time, 单位: 10小时), 每种变量的搭配用于随机选择的4个动物(共有48个观测). 从这个数据, 可以很容易算出在每个时间, 在不同条件下存活的动物比例, 这也就是对生存函数的估计. 类似地, 也可以估计危险函数. 由于数据的离散型, 这里估计出来的函数是阶梯状的. 图10.1是该数据对生存函数在各种情况下的估计的点图. 其中(a)为全部数据一起的图(虚线为置信带), (b)为按照不同的处理所做的图, (c)为按照不同毒药所做的图. 这些图的横坐标为生存的时间, 而纵坐标是生存函数的大小. 显然, 随着时间流逝, 生存的概率应该递减, 因此这种曲线都是呈下降趋势. 从图10.1可以看出不同处理和不同毒药对生存函数的影响.

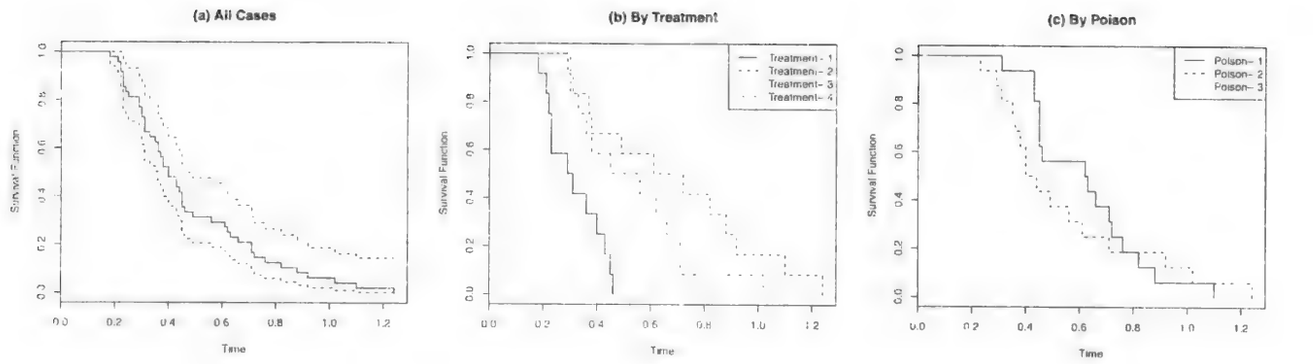

图 10.1 根据例10.1数据所产生的生存函数图, 其中(a)为全部数据一起的图(虚线为置信带), (b)为按照不同的处理所做的图, (c)为按照不同毒药所做的图. 可以看出不同处理和不同毒药对生存函数的影响.

图10.1是用程序包survival[1]的函数survfit()所绘, R代码(包括输入数据)如下:

```
w=read.table("poison.txt",header=T) #读入数据
library(survival);k=rep(1,48)
par(mfrow=c(1,3))
fit0=survfit(Surv(Time,k)~1,data=w)
plot(fit0,xlab="Time",ylab="Survival Function")
title("(a) All Cases")
fit1=survfit(Surv(Time,k)~Treatment,data=w)
plot(fit1,xlab="Time",ylab="Survival Function",lty=1:4)
title("(b) By Treatment")
legend("topright",paste("Treatment-",1:4),lty=1:4)
fit2=survfit(Surv(Time,k)~Poison,data=w)
plot(fit2,xlab="Time",ylab="Survival Function",lty=1:3)
title("(c) By Poison")
legend("topright",paste("Poison-",1:3),lty=1:3)
```

[1]Terry Therneau (2012). A Package for Survival Analysis in S. R package version 2.36-14.

一般来说, 数据中可能会有所谓的**删失的(censored)观测值**, 也就是说, 有些对象在实验过程中因为种种原因失去了记录(在某个时刻后消失). 以医药界的动物实验为例, 这种删失可能源于动物死于与实验无关的原因, 以医院病人为例, 病人可能自行出院, 无法跟踪调查等等. 例10.1没有删失, 下面看一个有删失观测值的数据.

例10.2 口咽癌数据(pharynx.txt, pharynx1.txt). 这个来自Kalbfleisch and Prentice (1980)的数据[1], 是基于美国Radiation Therapy Oncology Group的几个机构针对口咽若干位置的鳞状细胞癌的临床试验. 试验分成两组, 一组仅使用放疗(TX=1), 另一组放化疗皆用(TX=2). 原始数据为195 × 13的方阵. 这个数据是典型的生存分析数据, 可以用生存分析的方法, 比如Cox比例危险回归模型(Cox proportional hazards regression model), 也可以用其他回归方法. 下表是变量情况.

例10.2 口咽癌数据变量情况

变量名	描述	性质
CASE	编号	哑元型定性变量
INST	机构代码	哑元型定性变量
SEX	性别(1,2)	哑元型定性变量
TX	实验代码(1:标准,2:处理)	哑元型定性变量
GRADE	和正常细胞的区别度	哑元型定性变量
AGE	年龄	定量变量
COND	身体状况	哑元型定性变量
SITE	病变位置	哑元型定性变量
T.STAGE	癌症T分期	哑元型定性变量
N.STAGE	癌症N分期	哑元型定性变量
ENTRY.DT	进入试验日期	整数
STATUS	删失(0:右删失, 1:死亡)	哑元型定性变量
TIME	如未删失, 则是存活时间, 否则是最后记录时间(天数)	整数

读者可能发现, 这个数据中的CASE和ENTRY.DT不能参与建模, 应该删去. 在数据探索分析中还发现有两个观测值有缺失, 也予以删除. 此外, COND有些水平记录太少, 予以合并. 这样整理过的数据就是193 × 11的方阵(pharynx1.csv). 除了TIME和AGE之外都是分类(定性)变量. 变量TIME是这里关心的因变量.

生存数据都是按照一定格式记录的, 根据记录, 可以看出每个时间有多少死亡, 有多少存活, 各自的比例是多少等等, 这些记录被称为**生命表(Life Table)**. 下面对生命表予以介绍.

[1] Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*, John Wiley & Sons. 可从网站http://www.umass.edu/statdata/statdata/stat-nonlin.html 下载.

10.1 对生命数据的简单描述

生命表(Life Table)是对生存分析数据的一种数量和图形的描述. 通常, 生命表每一行代表一个时间段, 而该行的数据最少包括了在该时间段开始时有多少存活的对象, 有多少死亡的对象, 有多少删失的对象, 以及最基本的一些对生存函数和危险函数在该时间段的值的估计. 由于对于删失等观测值的处理方法不同, 对于简单的生命表有各种改进, 其中包括Kaplan-Meier方法、Fleming-Harrington方法等. 下面的表格是根据例10.2数据按照Kaplan-Meier方法所产生的生命表的前10行(一共180行). 这里一共三个表: 第一个是对照组和处理组混合(即全部)数据的生命表, 第二个是对照组(TX=1)的生命表, 第三个是处理组(TX=2)的生命表. 生命表的描述则为如图10.1那样的图形, 对于例10.2数据相应的三个图形在图10.2中. 但在前计算机时代, 这种图形则按照生命表手工画成.

例10.2分三种情况的生命表的前10行

混合数据					TX=1					TX=2				
t	r	e	c	s	t	r	e	c	s	t	r	e	c	s
11	193	2	0	0.99	81	98	1	0	0.99	11	95	2	0	0.98
15	191	1	0	0.98	89	97	1	0	0.98	15	93	1	0	0.97
38	190	1	0	0.98	90	96	0	1	0.98	38	92	1	0	0.96
74	189	1	0	0.97	94	95	1	0	0.97	74	91	1	0	0.95
81	188	1	0	0.97	99	94	2	0	0.95	105	90	1	0	0.94
89	187	1	0	0.96	112	92	2	0	0.93	107	89	1	0	0.93
90	186	0	1	0.96	127	90	1	0	0.92	112	88	1	0	0.92
94	185	1	0	0.96	128	89	1	0	0.91	130	87	1	0	0.91
99	184	2	0	0.95	144	88	1	0	0.90	134	86	1	0	0.89
105	182	1	0	0.94	147	87	1	0	0.89	147	85	1	0	0.88

上表中的t(time)代表时间, r(risk)代表在t时刻还没有死的人, e(event)代表那个时刻(时间段)发现死亡的人, c(censor) 代表那个时间段删失的人数s(survival function)代表利用某种方法(这里是Kaplan-Meier方法)计算的生存函数.

图10.2是由下面代码画出:

```
u=read.csv("pharynx1.txt",sep=",")#读入数据
library(survival)
fit0=survfit(Surv(TIME, STATUS) ~ 1, data=u ,type="kaplan-meier")
fit=survfit(Surv(TIME, STATUS) ~ TX, data=u ,type="kaplan-meier")
par(mfrow=c(1,2));
plot(fit0,con=F,xlab="Time",ylab="Survival Function")
title("All Data")
plot(fit,lty=1:2,xlab="Time",ylab="Survival Function")
title("Comparison");legend("topright",c("TX=1","TX=2"),lty=1:2)
```

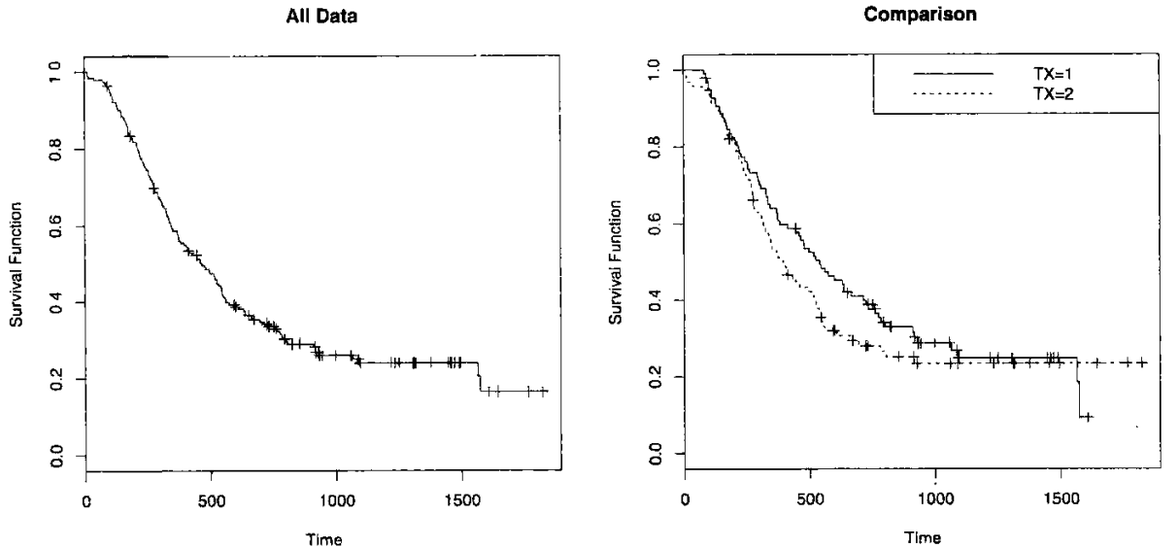

图 10.2　例10.2全部数据, 及比较TX=1和TX=2的生存函数图

这里不给出形成生命表的程序(放到后面小结中), 这是因为现在用不着再去还原过去作为原始数据的生命表, 生命表是数据分析的原始依据, 它仅仅给出了生存函数的估计. 从图10.2的右图可以看出, 化疗(TX=2)的生存函数在绝大部分时间中是在对照组的生存函数之下.

10.2　Cox 比例危险模型

回归的方法对于统计推断是十分重要的, 如何在生存数据的分析上建立类似于回归那样的模型呢? 人们一般希望生存函数能表示为某些相关的自变量的一个函数. 在例10.1中自变量就是处理和毒药, 例10.2中的自变量就是判别治疗组(TX=2)和对照组(TX=1)的哑元变量TX以及其他变量, 比如年龄、身体状况等等. 一般来说, 用x表示自变量(变量可能是向量, 即有多个自变量), 用$S(t|x)$表示在时间t的生存函数, 这里的x表示可能有关的自变量, 用$S_0(t)$表示待估计的基本生存函数(baseline survival function), 它和自变量x无关. Cox 比例危险模型为

$$S(t|x) = [S_0(t)]^{\exp(x^T\beta)}$$

这里的β为回归系数. 这里的线性部分$x^T\beta$是在$S_0(t)$的指数上面再取以e为底的指数. 当然该模型可以写成

$$\ln(-\ln S(t|x)) = x^T\beta + \ln H_0(t)$$

的线性形式. 这里基本累积危险函数$H_0(t)$是基本危险函数$h_0(t)$的积分

$$H_0 = \int_0^t h_0(u)du.$$

这里下标为0的, 名字冠以"基本"的$S_0(t), h_0(t)$和$H_0(t)$都是和x无关的. 细节可参见本章后面的公式. 注意只要得到$H(t), h(t)$和$S(t)$中之一, 就可以得到其他的.

$H_0(t), h_0(t)$和$S_0(t)$的关系也类似. 由此可以得到比例危险模型的其他形式:

$$\ln h(t|x) = \ln h_0(t) + x^T\beta$$

或者

$$h(t) = h_0(t)\exp(x^T\beta)$$

根据统计软件, 可以很容易得到对回归系数β的估计. 下面是用Cox比例危险模型拟合例16.1数据的代码, 其中最后一行代码是估计基本累积危险函数H_0(图形在图10.3中):

```
w=read.table("poison.txt",header=T)
for(i in 1:2)w[,i]=factor(w[,i])
k=rep(1,48);fit=coxph(Surv(Time,k)~.,data=w);summary(fit)
bh=basehaz(fit);plot(hazard~time,bh,type="l")
```

部分输出为:

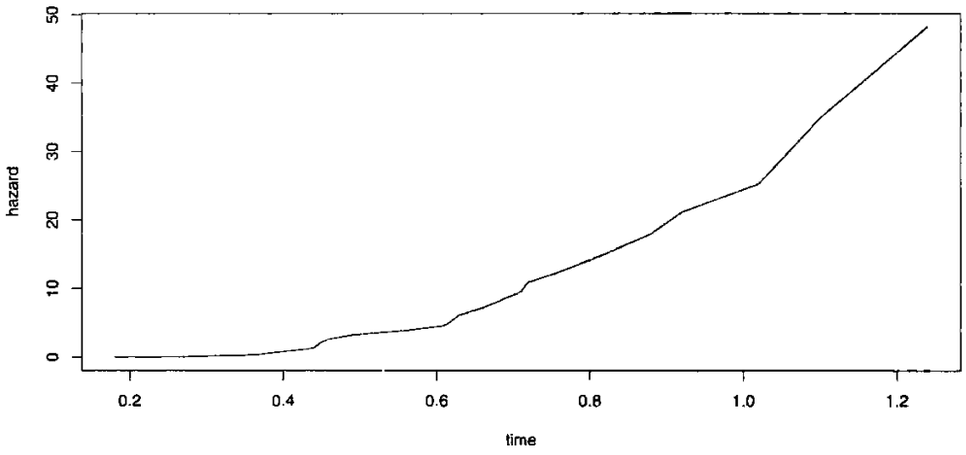

图 10.3 由例10.1根据Cox模型所估计的基本累积危险函数$H_0(t)$.

```
             coef exp(coef)  se(coef)      z Pr(>|z|)
Poison2   0.92862   2.53102   0.43310  2.144    0.032
Poison3   4.78337 119.50641   0.76088  6.287 3.25e-10
Treatment2 -3.52134   0.02956   0.60733 -5.798 6.71e-09
Treatment3 -1.17445   0.30899   0.45619 -2.574    0.010
Treatment4 -2.90340   0.05484   0.58878 -4.931 8.17e-07
```

如果令α_i $(i = 1,2,3)$表示毒药(Poison)的效应, 用β_j $(j = 1,2,3,4)$表示处理(treatment), 那么上面的输出表明我们的Cox模型为

$$\ln(-\ln S(t|x)) = \alpha_i + \beta_j + \ln H_0(t),\ i = 1,2,3,\ \ j = 1,2,3,4$$

这里$\alpha_1 = 0, \alpha_2 = 0.929, \alpha_3 = 4.783, \beta_1 = 0, \beta_2 = -3.521, \beta_3 = -1.174, \beta_4 = -2.903$. 一共有$3 \times 4 = 12$个方程.

10.3 小结

10.3.1 本章的概括和公式

1. 生存函数和危险函数

本章的基本函数为互相关联的生存函数$S(t)$、危险函数$h(t)$、累积危险函数$H(t)$(基本生存函数、基本危险函数和基本累积危险函数的关系类似). 生存函数定义为

$$S(t) = P(T \geq t) = 1 - F(T \leq t),\ t > 0,$$

这里T为作为随机变量的生存时间, $F(t)$为T的累积分布函数, 用$f(t)$表示其密度函数. 危险函数则定义为

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{[1 - F(t)]\Delta t}$$
$$= \frac{f(t)}{S(t)} = \frac{d}{dt}[-\ln S(t)].$$

累积危险函数为危险函数的积分

$$H(t) = \int_0^t h(u)du.$$

此外, 还有一些很容易推导出来的关系, 比如

$$H(t) = -\ln S(t);\ S(t) = \exp(-H(t)).$$

2. Cox 比例危险模型

用$S(t|x)$表示和自变量x有关的生存函数, 用$S_0(t)$表示与自变量无关的待估计基本生存函数. Cox比例危险模型为

$$S(t|x) = [S_0(t)]^{\exp(x^T\beta)}$$

这里的β为回归系数. 这里的线性部分$x^T\beta$是在$S_0(t)$的指数上面再取以e为底的指数. 当然该模型可以写成

$$\ln(-\ln S(t|x)) = x^T\beta + \ln H_0(t)$$

或者

$$h(t|x) = h_0(t)\exp(x^T\beta)$$

或其等价形式

$$\ln h(t|x) = \ln h_0(t) + x^T\beta.$$

之所以有"比例危险模型"这个词, 是因为对于不同的协变量x_i和x_j, 该模型满足危险函数的比例$h(t|x_i)/h(t|x_i)$与t无关. 类似于没有x的情况.

10.3.2 R语句的说明

使用程序包survival的Survfit()函数及coxph()函数时的因变量部分不是单纯的时间, 而是Surv(Time, status), 这里的Time是指生存时间, 而status(有时为censor, 依数据变量名字代码而定)为删失状况. Surv()函数要求删失为数值型(实际上是哑元定性变量)0,1变量, 0表示删失, 1表示没有删失.

在对例10.2的三个生命表(全部数据, TX=1和TX=2)的再现时, 使用了下面语句(在输入了数据之后):

```
library(survival)
f0=survfit(Surv(TIME,STATUS)~1,data=u ,type="kaplan-meier")
f1=survfit(Surv(TIME,STATUS)~1,data=u[u$TX==1,],type="kaplan-meier")
f2=survfit(Surv(TIME,STATUS)~1,data=u[u$TX==2,],type="kaplan-meier")
a0=cbind(t=f0$time,r=f0$n.risk,d=f0$n.event,c=f0$n.censor,s=f0$surv)
a1=cbind(t=f1$time,r=f1$n.risk,d=f1$n.event,c=f1$n.censor,s=f1$surv)
a2=cbind(t=f2$time,r=f2$n.risk,d=f2$n.event,c=f2$n.censor,s=f2$surv)
```

这里的a0, a1, a2就是所求的三个生命表. 其中type="kaplan-meier"意味着Kaplan-Meier方法, 也可以选择type="fleming-harrington".

10.4 习题

1. 重复对例10.1数据的计算, 改动一些选项, 尽可能解释输出图表的含义.

2. 用Cox比例危险模型拟合例10.2数据, 并解释结果.

第十一章　指数简介

11.1　指数漫谈

为了解一年来物价的总体变化, 有必要去了解每一项商品和服务的价格变化吗? 其实, 只要看一看公布的相关的**价格指数(price index)**就可以了. 这是因为计算物价指数的机构已经把不同时期的各种商品和服务(比如交通、娱乐、住房、食品和饮料、医疗、服装等)的价格和消费按照一些程序进行了调查. 他们把这些调查数据输入计算机, 根据某些公式进行计算, 并且和过去的某一标准进行对比. 然后在经过一些核对及调整之后公布对比结果, 也就是价格指数. 因此价格指数就是一种反映价格总体变化情况的综合变量.

从统计学的角度, **指数(index number)**就是代表所关心的变量的一些统计量. 在经济领域, 指数多为一些统计观测值的加权平均, 而且用过去类似的观测值平均作为基础, 以比例或百分比的形式出现. 上面说过的价格指数就有这样的形式. 综合指数并非经济学领域所专有的. 比如, 有衡量气象对人类或动物情绪、行为和生理影响的**生物气象指数(bioweather index)**, 有天文学家衡量星体颜色和温度的**颜色指数(color index)**, 有研究温度和湿度对人体舒适度影响的**温度湿度指数(Temperature-Humidity Index)**等等.

这些指数并不都是通过简单的算术或几何(加权)平均和比例而来的. 有些计算方法很复杂. 有些很简单. 方法也不全一样. 比如每个股票市场都有它自己的衡量股票价格的一些指数(虽然大同小异). 比如, **道琼斯指数、日经指数、恒生指数、纳斯达克指数、伦敦金融时报指数、上证综合指数、上证指数、深证综合指数、深圳成分股指数**等等. 为了可比性, 各国也采取一些同样(或类似)办法所计算的指数, 比如**国内生产总值(GDP)**等. 此外还有其他没有"指数"名称但也被认为是指数的统计量: 比如可以用来反映贫富差距的**Gini系数(Gini coefficient)**. 在我国的计划经济时期, 为了种种目的, 从苏联引进了大量的经济指数, 也自己编制了许多. 这些指数目前多数已经成为历史.

哪些统计量被称为指数, 并没有什么绝对限制, 依习惯而定. 当然, 不存在完美的指数. 即使对同一个对象和同一个目的, 可能会出现不同的指数: 只不过各自有各自的特点罢了. 许多指数的设计有很多不足之处, 但由于人们的习惯, 仍然在使用: 并没有把它们淘汰: 最多进行一些改进而已. 任何人都可能编制性质优秀的指数, 但是有没有人要用你编制的指数则是另外一件事了.

本章仅仅对一些常用指数及有关概念进行介绍. 这里不要求高等数学的知识, 但可能有一些不超过小学四则运算水平的简单公式.

11.2　价格指数

各个国家和地区都在编制自己的价格指数, 有些指数仅仅是针对部分产品而设计和计算的. 要想知道两个时期的价格的差距, 如何来计算呢? 你马上会说, 可

以用现在的价格除以过去的价格. 不错, 这就是价格指数的基础. 比如现在一公斤面粉是$P_t = 2$元, 去年是$P_0 = 1.6$元, 相对价格就是$P_1/P_0 = 2/1.6 = 1.25$. 为了去掉分数, 就乘以100, 得到(百分之)125. 但是, 单价并不代表你在面粉上花多少钱. 应该把你买了多少公斤面粉考虑进来. 但是用现在的购买量还是用过去的购买量计算, 就产生了不同的结果. 当然, 如果考虑食品价格, 就不能只考虑面粉一项. 假定作为比较基础的某年某商品的单价(或今年的单价)用P_0(或P_t)表示, 相应的消费量用Q_0(或Q_t)表示. 这里有四种计算总消费量的办法:

1. 各种商品的当年总消费为$\sum P_0Q_0$;

2. 按照今年的价格和当年的消费量的总额为$\sum P_tQ_0$;

3. 按照今年的价格和今年的消费量的总额为$\sum P_tQ_t$;

4. 按照当年的价格和今年的消费量的总额为$\sum P_0Q_t$.

看起来有些罗嗦, 但这是两种价格指数的计算基础. 一种称为**Laspeyres价格指数(Laspeyres price index)**, 另一种称为**Paasche价格指数(Paasche price index)**. 这些是Laspeyres(类)指数和Paasche(类)指数关于价格的形式; 这两个价格指数的定义分别为

$$\text{Laspeyres 价格指数} = \frac{\sum P_tQ_0}{\sum P_0Q_0}(100)$$

和

$$\text{Paasche 价格指数} = \frac{\sum P_tQ_t}{\sum P_0Q_t}(100).$$

它们的区别在于: 分子分母是全部使用过去的消费量Q_0, 还是全部使用目前的消费量Q_t. 很难从理论上说哪一个定义就一定比另一个优越, 但实际操作时可能有所不同. 显然, 对于Laspeyres价格指数, 作为计算不变的$\sum P_0Q_0$的基础年份就不能太特殊了, 需要有典型性. 作为这两个指数的几何平均的**Fisher理想指数(Fisher' s ideal index)**可以看成为这两个指数的折中方案. 它定义为

$$\text{Fisher理想指数} = \sqrt{(\text{Laspeyres 价格指数}) \times (\text{Paasche 价格指数})}.$$

11.3 数量指数(生活标准指数)

要想度量数量变化, 消费量在上面公式的分子中就一定要用Q_t, 而在分母中用Q_0. 但单价应该一样. 这种指数称为**数量指数(quantity index)**, 用来度量生活标准在量上的提高. 这时, 关于数量的Laspeyres指数和Paasche指数分别为

$$\text{Laspeyres 数量指数} = \frac{\sum P_0Q_t}{\sum P_0Q_0}(100)$$

和

$$\text{Paasche 数量指数} = \frac{\sum P_tQ_t}{\sum P_tQ_0}(100).$$

而Fisher理想指数仍然是这两个的几何平均.

11.4　总花费指数

要想得到总消费的变化, 分子的单价和消费量都应该是目前的, 而分母的单价和消费量都应该是作为基准的那一年的. 这样, Laspeyres指数、Paasche指数及Fisher理想指数就完全一样了, 统称为**总花费指数(total cost index)**. 显然

$$\text{总花费指数} = \frac{\sum P_t Q_t}{\sum P_0 Q_0}(100).$$

按照常理, 价格指数、数量指数及总花费指数应该满足下面关系:

$$\text{价格指数} \times \text{数量指数} = \text{总花费指数}.$$

但是在上面三个指数中, 如果利用这个乘积公式, Laspeyres指数过分估计总花费指数, 而Paasche指数又低估了它. 只有Fisher理想指数总是满足这个乘积关系. 因此才有"理想"的称号.

11.5　一两个常见的经济指数

这里并不想全面介绍各国都计算的所有指数. 下面仅仅介绍一两种新闻里出现较多的经济指数, 让大家有个感性认识.

1. **消费者价格指数**(consumer price index, CPI)

世界上有100多个国家都计算CPI. 虽然各个国家为计算CPI所使用的方法和覆盖的范围相差很大, 但总有很多共同的地方. 联合国每年都在其月度统计通报(Monthly Bulletin of Statistics)中公布各个国家的CPI.

在美国, 这是媒体中最经常出现的价格指数. 每个月经白宫认可由美国劳动统计局(Bureau of Labor Statistics)公布一次. 它是一个Laspeyres类型的指数. CPI抽取各种货物和服务的价格, 包括食品、房租和房价、能源、服装、交通、医药等. 每一个部分也都公布自己的指数. 这些部分按照重要性加权. 各个区域甚至城市也都有自己的CPI. 计算CPI的品种数量通常是250到450种之间. 对于小国家或贫穷国家, 品种数量常常只有100到150种.

美国的CPI只覆盖薪金收入者, 无论是在一个家庭还是单独生活都算. 而英国的CPI覆盖所有的家庭, 但不包括那些户主的收入超过某一界限的家庭, 也不包括那些至少四方之三的收入来自退休金的人. 很多国家在计算CPI时, 只考虑城市居民, 甚至少数城市. 比如澳大利亚只考虑各州首府, 墨西哥只考虑首都墨西哥城. 但有些则包括得广泛些. 比如日本包括了所有城乡家庭, 但不包括单人家庭和家长是农民和渔民的家庭. 由于这些限度, 为了更广泛的需要, 比如要度量国家福利的变化, 就需要包括所有人的更加复杂的指数, 比如包括单人家庭、乡村家庭和城市高收入家庭等等.

2. **批发价格指数**(wholesale price index)

中国目前还没有批发价格指数. 该指数度量制造者和批发商所给出的价格的变化. 它可能衡量到达零售商之前的一些有选择的阶段的货品价格变化. 它包括或者制造商对批发商所提出的价格, 或者批发商对零售商所提出的价格, 或者是这二者及其他中间人价格的组合.

在美国, 批发价格指数度量所有流入初级市场的国产或进口商品的价格变化. 初级市场是商品第一次以相当数量出售的市场. 商品在其各个加工阶段都有标价. 比如棉花在初级市场就有原棉、棉纱、棉布等各种价格形式. 批发价格指数在英美已经有一百多年的历史了.

批发价格指数所覆盖的商品数量在工业大国都有数千种, 而在多数国家常常只有一二百种. 如果只需要关于一般的总体商品的指数, 那么数量少些也够用了. 但如果需要许多分类子指数(subindex), 则需要包括很多的品种. 这些类别包括诸如初级产品、中间产品和最终产品, 或者耐用商品和不耐用商品等等. 在美国有15个范畴, 有接近100个子类(比如新鲜水果、谷物等)及大量的产品类(如苹果、香蕉、大麦玉米等). 对每一种范畴都有按月度公布的指数. 在工业不是那么多元化的国家, 类别的数量就要少些. 各个国家的批发价格指数都能够很好地代表原材料和标准产品, 而对于诸如重型电气设备的复杂产品则在先进的工业化国家代表不足甚至忽略. 这在总批发价格指数上造成一个向上的偏差, 因为有理由相信, 技术改进在改进复杂商品上是很重要的.

11.6 小结

本章简单介绍了指数的知识. 指数在统计上就是一些统计量. 它可以出现在任何领域. 在经济领域, 指数多为一些统计观测值的加权平均, 而且以过去类似观测值的平均作为基础的比例或百分比的形式出现. 我们既没有给出什么理论, 也没有给出习题. 由于各种指数是为其各自目的服务的, 一般由有关领域的权威、首脑或专家来确定. 实际上, 永远无法从理论上说明一种指数是绝对最优的. 现存的指数都有各种各样的毛病. 但是有多少人愿意放弃他们熟悉而又实用(虽然有不足)的事物而去采用一些陌生的新事物呢?

附录A　练习:熟练使用R软件

实践1(最初几步):

```
x=1:100#把1,2,...,100个整数向量赋值到x
(x=1:100) #同上，只不过显示出来
sample(x,20) #从1,...,100中随机不放回地抽取20个值作为样本
set.seed(0);sample(1:10,3)#先设随机种子再抽样.
#从1,...,200000中随机不放回地抽取10000个值作为样本:
z=sample(1:200000,10000)
z[1:10]#方括号中为向量z的下标
y=c(1,3,7,3,4,2)
z[y]#以y为下标的z的元素值
(z=sample(x,100,rep=T))#从x放回地随机抽取100个值作为样本
(z1=unique(z))
length(z1)#z中不同的元素个数
xz=setdiff(x,z) #x和z之间的不同元素--集合差
sort(union(xz,z))#对xz及z的并的元素从小到大排序
setequal(union(xz,z),x) #对xz及z的并的元素与x是否一样
intersect(1:10,7:50) #两个数据的交
sample(1:100,20,prob=1:100)#从1:100中不等概率随机抽样,
        #各数目抽到的概率与1:100成比例
```

实践2(一些简单运算):

```
pi *10^2 #能够用?" *" 来看基本算术运算方法，pi是圆周率
"*"(pi, "^"(10,2)) #和上面一样，有些繁琐，是吧！没有人这么用
pi * (1:10)^-2.3#可以对向量求指数幂
x = pi * 10^2
x
print(x) #和上面一样
(x=pi *10^2) #赋值带打印
pi^(1:5) #指数也可以是向量
print(x, digits = 12)#输出x的12位数字
```

实践3(关于R对象的类型等):

```
x=pi*10^2
class(x) #x的class
typeof(x) #x的type
class(cars)#cars是一个R中自带的数据
typeof(cars) #cars的type
names(cars)#cars数据的变量名字
```

```
summary(cars) #cars的汇总
head(cars)#cars的头几行数据, 和cars[1:6,]相同
tail(cars) #cars的最后几行数据
str(cars)#也是汇总
row.names(cars) #行名字
attributes(cars)#cars的一些信息
class(dist~speed)#公式形式,"~"左边是因变量,右边是自变量
plot(dist ~speed,cars)#两个变量的散点图
plot(cars$speed,cars$dist) #同上
```

实践4(包括简单自变量为定量变量及定性变量的回归):

```
ncol(cars);nrow(cars) #cars的行列数
dim(cars) #cars的维数
lm(dist ~ speed, data = cars)#以dist为因变量,speed为自变量做OLS
cars$qspeed =cut(cars$speed, breaks=quantile(cars$speed),
    include.lowest = TRUE) #增加定性变量qspeed, 四分位点为分割点
names(cars) #数据cars多了一个变量
cars[3]#第三个变量的值和cars[,3]类似
table(cars[3])#列表
is.factor(cars$qspeed)
plot(dist ~ qspeed, data = cars)#点出箱线图
(a=lm(dist ~ qspeed, data = cars))#拟合线性模型(简单最小二乘回归)
summary(a)#回归结果(包括一些检验)
```

实践5(简单样本描述统计量等等):

```
x <- round(runif(20,0,20), digits=2)#四舍五入
summary(x) #汇总
min(x);max(x) #极值, 与range(x)类似
median(x)  # 中位数(median)
mean(x)    # 均值(mean)
var(x)     #方差(variance)
sd(x)      # 标准差(standard deviation),为方差的平方根
sqrt(var(x)) #平方根
rank(x)    # 秩(rank)
order(x)#升幂排列的x的下标
order(x,decreasing = T)#降幂排列的x的下标
x[order(x)] #和sort(x)相同
sort(x)     #同上: 升幂排列的x
sort(x,decreasing=T)#sort(x,dec=T) 降幂排列的x
sum(x);length(x)#元素和及向量元素个数
```

```
round(x) #四舍五入,等于round(x,0),而round(x,5)为留到小数点后5位
fivenum(x)  # 五数汇总, quantiles
quantile(x) # 分位点 quantiles (different convention)有多种定义
quantile(x, c(0,.33,.66,1))
mad(x) # "median average distance":
cummax(x)#累积最大值
cummin(x)#累积最小值
cumprod(x)#累积积
cor(x,sin(x/20)) #线性相关系数 (correlation)
```

实践6(简单图形):

```
x=rnorm(200)#200个随机正态数赋值到x
hist(x, col = "light blue")#直方图(histogram)
rug(x) #在直方图下面加上实际点的大小
stem(x)#茎叶图
x <- rnorm(500)
y <- x + rnorm(500) #构造一个线性关系
plot(y~ x) #散点图
a=lm(y~x) #做回归
abline(a,col="red")#或者abline(lm(y~x),col="red")散点图加拟合线
print("Hello World!")
paste("x 的最小值= ", min(x)) #打印
demo(graphics)#演示画图(点Enter来切换)
```

实践7(复数运算和求函数极值):

```
(2+4i)^-3.5+(2i+4.5)*(-1.7-2.3i)/((2.6-7i)*(-4+5.1i))#复数运算
#下面构造一个10维复向量, 实部和虚部均为10个标准状态样本点:
(z <-complex(real=rnorm(10), imaginary =rnorm(10)))
complex(re=rnorm(3),im=rnorm(3))#3维复向量
Re(z) #实部
Im(z) #虚部
Mod(z) #模
Arg(z) #辐角
choose(3,2) #组合
factorial(6)#排列6!
#解方程:
f=function(x) x^3-2*x-1
uniroot(f,c(0,2))#迭代求根
#如果知道根为极值
f=function(x) x^2+2*x+1 #定义一个二次函数
```

```
optimize(f,c(-2,2))#在区间(-2,2)间求极值
```

实践8(字符型向量):

```
a=factor(letters[1:10])#letters:小写字母的向量,LETTERS:大写字母
a[3]="w"              #不行! 会给出警告
a=as.character(a) #转换一下
a[3]="w"              #可以了
a;factor(a)           #两种不同的类型
```

实践9(数据输入输出):

```
x=scan()#从屏幕输入数据, 可以键入, 也可以粘贴,可多行输入,空行后Enter
1.5 2.6 3.7 2.1 8.9 12 -1.2 -4

x=c(1.5,2.6,3.7,2.1,8.9,12,-1.2,-4)#等价于上面
w=read.table(file.choose(),header=T)#从列表中选择有变量名的数据
setwd("f:/2010stat")#或setwd("f:\2010stat")#建立工作路径
(x=rnorm(20)) #给x赋值20个标准正态数据值
#(注:有常见分布的随机数, 分布函数,密度函数及分位数函数)
write(x,"f:/2010stat/test.txt")#把数据写入文件(路径要对)
y=scan("f:/2010stat/test.txt");y #扫描文件数值数据到y
y=iris;y[1:5,];str(y) #iris是R自带数据
write.table(y,"test.txt",row.names=F)#把数据写入文本文件
w=read.table("f:/2010stat/test.txt",header=T)#读带有变量名的数据
str(w) #汇总
write.csv(y,"test.csv")#把数据写入csv文件
v=read.csv("f:/2010stat/test.csv")#读入csv数据文件
str(v) #汇总
data=read.table("clipboard")#读入剪贴板的数据
```

实践10(序列等等):

```
(z=seq(-1,10,length=100))#-1到10等间隔的100个数的序列
z=seq(-1,10,len=100)#和上面等价写法
(z=seq(10,-1,-0.1)) #10到-1间隔为-0.1的序列
(x=rep(1:3,3))  #三次重复1:3
(x=rep(3:5,1:3)) #自己看, 这又是什么呢?
x=rep(c(1,10),c(4,5))
w=c(1,3,x,z);w[3]#把数据(包括向量)组合(combine)成一个向量
x=rep(0,10);z=1:3;x+z #向量加法(如果长度不同, R如何给出警告和结果?)
x*z    #向量乘法
rev(x)#颠倒次序
z=c("no cat","has ","nine","tails") #字符向量
```

```
z[1]=="no cat" #双等号为逻辑等式
z=1:5
z[7]=8;z #什么结果? 注:NA为缺失值(not available)
z=NULL
z[c(1,3,5)]=1:3;
z
rnorm(10)[c(2,5)]
z[-c(1,3)]#去掉第1、3元素
z=sample(1:100,10);z
which(z==max(z))#给出最大值的下标
```

实践11(矩阵):

```
x=sample(1:100,12);x #抽样
all(x>0);all(x!=0);any(x>0);(1:10)[x>0]#逻辑符号的应用
diff(x) #差分
diff(x,lag=2) #差分
x=matrix(1:20,4,5);x #矩阵的构造
x=matrix(1:20,4,5,byrow=T);x#矩阵的构造, 按行排列
t(x) #矩阵转置
x=matrix(sample(1:100,20),4,5)
2*x
x+5
y=matrix(sample(1:100,20),5,4)
x+t(y) #矩阵之间相加
(z=x%*%y) #矩阵乘法
z1=solve(z) # solve(a,b)可以解ax=b方程
z1%*%z #应该是单位向量, 但浮点运算不可能得到干净的0
round(z1%*%z,14)  #四舍五入
b=solve(z,1:4); b #解联立方程
```

实践12(矩阵继续):

```
nrow(x);ncol(x);dim(x)#行列数目
x=matrix(rnorm(24),4,6)
x[c(2,1),]#第2和第1行
x[,c(1,3)] #第1和第3列
x[2,1] #第[2,1]元素
x[x[,1]>0,1] #第1列大于0的元素
sum(x[,1]>0) #第1列大于0的元素的个数
sum(x[,1]<=0) #第1列不大于0的元素的个数
x[,-c(1,3)]#没有第1、3列的x.
```

```
diag(x)  #x的对角线元素
diag(1:5) #以1:5为对角线,其他元素为0的对角线矩阵
diag(5) #5维单位矩阵
x[-2,-c(1,3)]#没有第2行、第1、3列的x
x[x[,1]>0&x[,3]<=1,1]#第1列>0并且第3列<=1的第1列元素
x[x[,2]>0|x[,1]<.51,1]#第1列<.51或者第2列>0的第1列元素
x[!x[,2]<.51,1]#第1列中相应于第2列中>=.51的元素
apply(x,1,mean)#对行(第一维)求均值
apply(x,2,sum)#对列(第二维)求和
x=matrix(rnorm(24),4,6)
x[lower.tri(x)]=0;x #得到上三角阵,
#为得到下三角阵, 用x[upper.tri(x)]=0)
```

实践13(高维数组):

```
x=array(runif(24),c(4,3,2))
x#从24个均匀分布的样本点构造4乘3乘2的三维数组
is.matrix(x)
dim(x)#得到维数(4,3,2)
is.matrix(x[1,,])#部分三维数组是矩阵
x=array(1:24,c(4,3,2))
x[c(1,3),,]
x=array(1:24,c(4,3,2))
apply(x,1,mean)    #可以对部分维做均值运算
apply(x,1:2,sum)   #可以对部分维做求和运算
apply(x,c(1,3),prod) #可以对部分维做求乘积运算
```

实践14(矩阵与向量之间的运算):

```
x=matrix(1:20,5,4) #5乘4矩阵
sweep(x,1,1:5,"*")#把向量1:5的每个元素乘到每一行
sweep(x,2,1:4,"+")#把向量1:4的每个元素加到每一列
x*1:5
%sweep(x,2,1:4,"+")#标准化,即每一元素减去该列均值,除以该列标准差:
(x=matrix(sample(1:100,24),6,4));(x1=scale(x))
(x2=scale(x,scale=F))#自己观察并总结结果
(x3=scale(x,center=F)) #自己观察并总结结果
round(apply(x1,2,mean),14) #自己观察并总结结果
apply(x1,2,sd)#自己观察并总结结果
round(apply(x2,2,mean),14);apply(x2,2,sd)#自己观察并总结结果
round(apply(x3,2,mean),14);apply(x3,2,sd)#自己观察并总结结果
```

实践15(缺失值, 数据的合并):

```
airquality #有缺失值(NA)的R自带数据
complete.cases(airquality)#判断每行有没有缺失值
which(complete.cases(airquality)==F) #有缺失值的行号
sum(complete.cases(airquality)) #完整观测值的个数
na.omit(airquality) #删去缺失值的数据
#附加, 横或竖合并数据: append,cbind,rbind
x=1:10;x[12]=3
(x1=append(x,77,after=5))
cbind(1:5,rnorm(5))
rbind(1:5,rnorm(5))
cbind(1:3,4:6);rbind(1:3,4:6) #去掉矩阵重复的行
(x=rbind(1:5,runif(5),runif(5),1:5,7:11))
x[!duplicated(x),]
unique(x)
```

实践16(list):

```
#list可以是任何对象(包括list本身)的集合
z=list(1:3,Tom=c(1:2,a=list("R",letters[1:5]),w="hi!"))
z[[1]];z[[2]]
z$T
z$T$a2
z$T[[3]]
z$T$w
```

实践17(条形图和表):

```
x =scan()#30个顾客在五个品牌中的挑选
3 3 3 4 1 4 2 1 3 2 5 3 1 2 5 2 3 4 2 2 5 3 1 4 2 2 4 3 5 2

barplot(x) #不合题意的图
table(x) #制表
barplot(table(x)) #正确的图
barplot(table(x)/length(x)) #比例图(和上图形状一样)
table(x)/length(x)
```

实践18(形成表格):

```
library(MASS)#载入软件包MASS
quine #MASS所带数据
attach(quine)#把数据变量的名字放入内存
#下面是从该数据得到的各种表格
table(Age)
table(Sex, Age); tab=xtabs(~ Sex + Age, quine); unclass(tab)
```

```
tapply(Days, Age, mean)
tapply(Days, list(Sex, Age), mean)
detach(quine) #attach的逆运行
```

实践19(如何写函数):

```
#下面这个函数是按照定义(编程简单, 但效率不高)求n以内的素数
ss=function(n=100){z=2;
for (i in 2:n)if(any(i%%2:(i-1)==0)==F)z=c(z,i);return(z) }
fix(ss) #用来修改任何函数或编写一个新函数
ss() #计算100以内的素数
t1=Sys.time() #记录时间点
ss(10000) #计算10000以内的素数
Sys.time()-t1 #费了多少时间
system.time(ss(10000))#计算执行ss(10000)所用时间
#函数可以不写return,这时最后一个值为return的值.
#为了输出多个值最好使用list输出
```

实践20(画图):

```
x=seq(-3,3,len=20);y=dnorm(x)#产生数据
w= data.frame(x,y)#合并x,成为数据w
par(mfcol=c(2,2))#准备画四个图的地方
plot(y ~ x, w,main="正态密度函数")
plot(y ~ x,w,type="l", main="正态密度函数")
plot(y ~ x,w,type="o", main="正态密度函数")
plot(y ~ x,w,type="b",main="正态密度函数")
par(mfcol=c(1,1))#取消par(mfcol=c(2,2))
```

实践21(色彩和符号等调节):

```
plot(1,1,xlim=c(1,7.5),ylim=c(0,5),type="n") #画出框架
#在plot命令后面追加点(如要追加线可用lines函数):
points(1:7,rep(4.5,7),cex=seq(1,4,l=7),col=1:7, pch=0:6)
text(1:7,rep(3.5,7),labels=paste(0:6,letters[1:7]),cex=seq(1,4,l=7),
col=1:7)#在指定位置加文字
points(1:7,rep(2,7), pch=(0:6)+7)#点出符号7到13
text((1:7)+0.25, rep(2,7), paste((0:6)+7))#加符号号码
points(1:7,rep(1,7), pch=(0:6)+14) #点出符号14到20
text((1:7)+0.25, rep(1,7), paste((0:6)+14)) #加符号号码
#这些关于符号形状、大小、颜色以及其他画图选项的说明可用"?par"来查看
```